HUMBOLDT-UNIVERSITÄT ZU BERLIN
INSTITUT FÜR BIBLIOTHEKS- UND INFORMATIONSWISSENSCHAFT

BERLINER HANDREICHUNGEN
ZUR BIBLIOTHEKS- UND
INFORMATIONSWISSENSCHAFT

HEFT 329

AUTOMATISCHE INDEXIERUNG IN DER SOZIALWISSENSCHAFTLICHEN FACHINFORMATION

EINE EVALUATIONSSTUDIE ZUR MASCHINELLEN ERSCHLIEßUNG FÜR DIE DATENBANK SOLIS

VON
ANDREAS OSKAR KEMPF

Automatische Indexierung in der Sozialwissenschaftlichen Fachinformation

Eine Evaluationsstudie zur maschinellen Erschließung für die Datenbank SOLIS

von
Andreas Oskar Kempf

Berliner Handreichungen zur
Bibliotheks- und Informationswissenschaft

Begründet von Peter Zahn
Herausgegeben von
Konrad Umlauf
Humboldt-Universität zu Berlin

Heft 329

Kempf, Andreas Oskar

Automatische Indexierung in der sozialwissenschaftlichen Fachinformation : Eine Evaluationsstudie zur maschinellen Erschließung für die Datenbank SOLIS / von Andreas Oskar Kempf. - Berlin : Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin, 2012. - 127 S. : graph. Darst. - (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft ; 329)

ISSN 14 38-76 62

Abstract:

Automatische Indexierungsverfahren werden mit Zunahme der digitalen Verfügbarkeit von Metadaten und Volltexten mehr und mehr als eine mögliche Antwort auf das Management unstrukturierter Daten diskutiert. In der sozialwissenschaftlichen Fachinformation existiert in diesem Zusammenhang seit einiger Zeit der Vorschlag eines sogenannten Schalenmodells (vgl. Krause, 1996) mit unterschiedlichen Qualitätsstufen bei der inhaltlichen Erschließung. Vor diesem Hintergrund beschreibt die Arbeit zunächst Methoden und Verfahren der inhaltlichen und automatischen Indexierung, bevor vier Testläufe eines automatischen Indexierungssystems (MindServer) zur automatischen Erschließung von Datensätzen der bibliographischen Literaturdatenbank SOLIS mit Deskriptoren des Thesaurus Sozialwissenschaften sowie der Klassifikation Sozialwissenschaften beschrieben und analysiert werden. Es erfolgt eine ausführliche Fehleranalyse mit Beispielen sowie eine abschließende Diskussion, inwieweit die automatische Erschließung in dieser Form für die Randbereiche der Datenbank SOLIS für die Zukunft einen gangbaren Weg darstellt.
Diese Veröffentlichung geht zurück auf eine Masterarbeit im postgradualen Fernstudiengang Master of Arts (Library and Information Science) an der Humboldt-Universität zu Berlin.

Online-Version: http://edoc.hu-berlin.de/series/berliner-handreichungen/2012-329

Dank

Diese Arbeit behandelt ein komplexes Themenfeld, das in der zur Verfügung stehenden Zeit nur mit besonderer Unterstützung bearbeitet werden konnte. Ich möchte mich daher bei zahlreichen Personen bedanken, die zur Entstehung dieser Arbeit beigetragen haben.

In besonderer Weise gilt mein Dank den Mitarbeiter-/innen bei GESIS | Leibniz-Institut für Sozialwissenschaften, die mich bei dem Aufbau und bei der Durchführung sowie zum Teil bei der Auswertung der Testläufe unterstützt haben, namentlich Monika Zimmer, Hannelore Schott und Jan Hendrik Schulz.[1] Bei Philipp Schaer und Nadine Dulisch möchte ich mich vor allem für die technische Unterstützung bedanken.

Bedanken möchte ich mich auch bei meinen beiden Betreuer/-innen. Bei Vivien Petras für die intensive Betreuung und den kontinuierlichen Dialog über den gesamten Zeitraum der Arbeit. Aus der räumlichen Distanz ging eine besondere Intensität in der Betreuung hervor. Bei Stefan Gradmann bedanke ich mich vor allem dafür, dass er derart spontan die Zweitbetreuung zugesagt hat.

Daneben möchte ich mich bei den Veranstalter-/innen des PETRUS-Workshops an der Deutschen Nationalbibliothek in Frankfurt am Main, Christa Schöning-Walter und Reinhard Altenhöner bedanken. Nicht nur hatte ich durch den Einblick in das Projekt PETRUS an der Deutschen Nationalbibliothek im Rahmen eines Praktikums erstmals die Möglichkeit, mit automatischer Indexierung in Berührung zu kommen. Der PETRUS-Workshop, ein knappes Jahr später, gab mir zusätzlich die Gelegenheit, mein Vorgehen bei der Durchführung der Testläufe in einem Kreis von Fachkolleg-/innen vorzustellen und zu diskutieren.

[1] Bei der Ansprache konkreter Personenkreise wird eine Gender gerechte Sprache verwendet. Ansonsten wird aus Gründen der Lesbarkeit darauf verzichtet. Gleichwohl sind beide Geschlechter einbezogen.

Inhalt

Einleitung

Über wissenschaftliche Informationen zu verfügen und auf diese unter geringem Ressourcenaufwand gezielt zuzugreifen, ist von hoher gesellschaftlicher Relevanz. Eine Einsicht, die etwa im Zuge des sogenannten Sputnik-Schocks in den 1960er und 70er Jahren den Grundstein für die Bildung der Fachinformationszentren in Deutschland legte. Längst kursiert für die postindustrielle Gesellschaft der Begriff der Informationsgesellschaft, um dem explosionsartigen Anstieg verfügbarer Informationen durch Datennetze und der wachsenden wirtschaftlichen Bedeutung des Informationssektors Rechnung zu tragen (vgl. Castells 2001 sowie Steinbicker 2001). Dabei lassen sich im öffentlichen Sektor, parallel zum Anstieg der Publikationstätigkeit, eine Stagnation bzw. ein Rückgang der Ressourcen für die Informationserschließung beobachten.

Mit Zunahme der digitalen Verfügbarkeit von Metadaten und Volltexten werden automatische Indexierungsverfahren als eine mögliche Antwort auf das Management unstrukturierter Daten diskutiert.[2] Dabei reichen die Anfänge wissenschaftlicher Forschung zur maschinellen Unterstützung des Indexierungsvorgangs bis in die 1960er Jahre zurück. Vor allem in einem klassischen Anwendungsbereich der dokumentarischen Profession, der Medien- bzw. Pressedokumentation, wurde früh mit der automatischen Indexierung begonnen. Auf der Grundlage schmaler Thesauri und Klassifikationen lieferten die angewendeten Verfahren schnelle Erfolge und eine deutliche Kosten- und Zeitersparnis.[3]

[2] In der vorliegenden Arbeit werden die Bezeichnungen „automatisches Indexieren“ und „maschinelles Indexieren“ synonym verwendet. In Bezug auf den konkreten Anwendungsfall, der inhaltlichen Erschließung in der sozialwissenschaftlichen Fachinformation, wird dabei allerdings stets von einem semiautomatischen bzw. prozessunterstützenden Erschließungsverfahren ausgegangen. Eine solche Differenzierung zwischen (voll)automatischen und semiautomatischen Verfahren steht dabei nicht so sehr für verschiedenartiger Verfahrenssysteme als vielmehr für die Implementierung unterschiedlicher Arbeits- und Prozessabläufe.

[3] Beispiele bilden die Nachrichtenagentur Reuters und der Verlag Gruner + Jahr (vgl. Bertram 2005). Doch auch im Nachrichten- und Pressewesen gibt es umfangreiche Kategorienschemata. So reduzierte das ZDF im Jahr 2000, als es sich zur semiautomatischen Erschließung von Zeitungen und Zeitschriftenartikeln für die Indexierungssoftware Recommind entschied, den hauseigenen Thesaurus von über 3.000 auf knapp 2.000 Deskriptoren (vgl. Lingelbach-Hupfauer 2011).

Mit dem Umstieg auf Online-Kataloge in den 1990er Jahren setzte auch in den Bibliotheken eine Diskussion um die bisherige Form der Inhaltserschließung ein. Es stellte sich die Frage, inwiefern sich auch der Vorgang der Erschließung automatisieren ließe. Wichtige Initiativen zur Entwicklung und Anwendung computerunterstützender Inhaltserschließung, die zum Teil auf linguistischen, zum Teil auf statistischen Verfahren beruhen, bilden in der deutschsprachigen Bibliothekswelt die Projekte MILOS I und II, das Nachfolgeprojekt KASKADE und das Projekt OSIRIS (vgl. Siegmüller 2007). Daneben arbeitet die Deutsche Nationalbibliothek aktuell im Rahmen ihres Projekts PETRUS an der Entwicklung eines geeigneten computerunterstützenden Erschließungsverfahrens für Netzpublikationen (vgl. Schöning-Walter 2011 sowie DNB 2010).

Seit etwa zehn Jahren sammeln auch eine Reihe von Fachinformationszentren Erfahrungen auf dem Gebiet der automatischen Erschließung. Mit dem Auftrag, die jeweilige Fachgemeinschaft mit relevanter wissenschaftlicher Information zu versorgen, obliegt ihnen in besonderem Maße eine zeitnahe Strukturierung und tiefgehende Aufbereitung von Informationen. Gleichzeitig zeichnet sich insgesamt eine Veränderung in den Recherchegewohnheiten der Nutzer von Informationsangeboten ab. Die Verwendung fachspezifischer kontrollierter Vokabulare und geschickter Suchstrategien wird zunehmend durch die einfache Freitextsuche ersetzt (vgl. ITA 2010 sowie Stahl et al. 2005).

Bereits früh sammelte die Zentralbibliothek für Wirtschaftswissenschaften (ZBW) mit dem von der Deutschen Forschungsgemeinschaft (DFG) geförderten Projekt AUTINDEX (AUTomatische INDEXierung) Erfahrungen auf dem Gebiet der automatischen Indexierung. Unter Anwendung der gleichnamigen, vornehmlich computerlinguistisch operierenden Indexierungssoftware des Instituts für Angewandte Informationsforschung (IAI) ging daraus im Jahr 2004 ein Prototyp zur Implementierung eines semiautomatischen Indexierungsverfahrens hervor, der allerdings nicht eingesetzt werden konnte, da sich zu diesem Zeitpunkt keine Verbindung der Indexierungssoftware mit dem Produktionssystem herstellen ließ. Seit dem Jahr 2009 besteht ein neues Projekt zur Vorbereitung einer automatischen Erschließung auf der Grundla-

ge der in erster Linie statistischen Softwareanwendung MindServer. Durch die ausschließliche Zuordnung von Deskriptoren aus dem Standard-Thesaurus Wirtschaft wird der Erschließungsvorgang um einen begriffsorientierten Verfahrensansatz ergänzt.[4] (Zur Unterscheidung der verschiedenen Verfahrensansätze siehe Teil II).

Ein Beispiel für eine bereits implementierte prozessunterstützende Erschließung auf der Grundlage der Software AUTINDEX bildet die Dokumentation psychologischer Literatur und Medien in der Datenbank PSYNDEX am Zentrum für Psychologische Information und Dokumentation (ZPID).[5] Die automatische Indexierungssoftware wurde in den Dokumentationsablauf integriert und liefert Deskriptorenvorschläge zur Unterstützung der inhaltlichen Erschließung durch manuelle Indexierer. Einen zentralen Zwischenschritt bei der Einführung des semiautomatischen Indexierungsverfahrens bildete dabei der Aufbau einer Indikatorenliste. Durch sie wurde der Thesaurus der *American Psychological Association* (APA) um 23.500 Begriffe ergänzt, um die Repräsentation der Dokumentinhalte durch geeignete Deskriptoren zu erleichtern.[6] Grundlage der Indexierungssoftware bilden Abstracts, Dokumenttitel sowie von den Autoren angegebene Schlagwörter.[7]

Mit dem Ziel, durch ein semiautomatisches Indexierungsverfahren Arbeitszeit einzusparen, beschäftigt sich seit einiger Zeit auch GESIS │Leibniz-Institut

[4] Aus einer Evaluationsstudie liegen bereits Ergebnisse zum Indexierungsverhalten der Software vor. Vergleichen mit der manuellen Indexierung wurde eine Übereinstimmung der Erschließungsergebnisse des automatischen Verfahrens mit den intellektuell erstellten Indexierungsvorgaben zu 36 Prozent erzielt (vgl. Groß 2010).

[5] Im Kern beruht die Software auf einer natürlichsprachigen morpholinguistischen Textanalyse. In Verbindung mit einer statistischen Ableitung semantischer Klassen gehen hieraus Deskriptorenvorschläge hervor, die mit dem kontrollierten Vokabular des Thesaurus abgeglichen werden (vgl. Gerards/Gerards/Weiland 2006).

[6] Eine Evaluation der Indexierungssoftware anhand einer Stichprobe, bei der die manuellen Indexierungsergebnisse als Vergleichsdaten herangezogen wurden, ergab eine durchschnittliche Übereinstimmung der automatisch generierten Deskriptorenvorschläge mit der manuellen Indexierung um 46,7 Prozent (*Recall*). Bezogen auf die durchschnittliche Gesamtzahl der automatisch vergebenen Deskriptoren bedeutete dies eine Übereinstimmung mit dem intellektuell generierten Indexat um 35,4 Prozent (*Precision*) (vgl. ebd. sowie Gerards 2011).

[7] Seit kurzem sammelt auch das Deutsche Institut für Internationale Pädagogische Forschung (DIPF) Erfahrungen auf dem Gebiet der automatischen Indexierung. Nach einer Marktanalyse und dem Einholen von Erfahrungsberichten liegen mittlerweile Ergebnisse zu den Tests zweier Softwareprodukte, sowohl eines linguistischen als auch eines statistischen Verfahrens, vor. Die Berechnung der Indexierungskonsistenz als Maß für die Übereinstimmung zwischen unterschiedlichen Indexaten derselben Vorlage – in diesem Fall von Indexaten, die im einen Fall auf einer intellektuellen und im anderen Fall auf einer automatischen Indexierung beruhen – ergab Werte von jeweils knapp unter 30 Prozent (vgl. Wissel 2011).

für Sozialwissenschaften mit dem Thema der softwareunterstützten Erschließung. Den Ausschlag dafür gab zum einen die Anschaffung der Software MindServer zur Entwicklung eines *Search Term Recommenders* für das Fachportal sowiport im Rahmen eines des DFG-geförderten Projekts zu Mehrwertdiensten für das Information Retrieval (IRM) (vgl. Mayr et al. 2009).[8] Zum anderen zeichneten sich für die Fachabteilung zur Produktion und Pflege der Datenbanken und Informationssysteme neue Aufgabenbereiche ab, die mittelfristig die Freisetzung von Ressourcen erforderlich erscheinen ließen.

Erste Gedanken zur Weiterentwicklung der inhaltlichen Erschließung reichen in der sozialwissenschaftlichen Fachinformation allerdings noch weiter zurück. So brachte der frühere Präsident des Instituts, Jürgen Krause (1996, 2006), die Vorstellung eines Schalenmodells zur Inhaltserschließung in die Diskussion ein. Im Zuge der verstärkten Deregulation im Informationswesen durch den Rückzug des Staats aus dem Informationsmarkt tritt er für neue Strukturen in der Inhaltserschließung ein. Integriert in ein Informationssystem, das aus den Datenlieferungen unterschiedlicher Partner hervorgehe, sei je nach Datenrelevanz die inhaltliche Erschließung nach unterschiedlichen Niveaustufen denkbar. Diese bildeten gleichsam verschiedene Schalen um einen tief und qualitativ hochwertig erschlossenen Kernbereich. Diese Überlegungen werden exemplarisch auf die Literaturdatenbank SOLIS bezogen. Demnach bildeten weniger relevant erscheinende Datenbestände, die zwar nach dem Fachthesaurus erschlossen allerdings ohne Abstract angeboten würden, ebenso eine der äußeren Schalen wie Zulieferungen, die mit einem anderen Vokabular erschlossen würden. Die in den Ausführungen äußerste Schale schlägt Krause für eine automatische Indexierung vor.

[8] Neben der Entwicklung einer Suchraum-Erweiterung ging hieraus ein Re-Ranking-Dienst auf Basis von Bradfordizing hervor (vgl. ebd. sowie Mayr 2010). Die Integration der Datenbank SOLIS in das sozialwissenschaftliche Fachportal *sowiport* bedeutet in diesem Zusammenhang für das Information Retrieval beispielsweise, dass eine Suchterm-Erweiterung eingebaut ist, die eine Unterstützung bei der Formulierung und Präzisierung der Suchanfrage bietet.

Zentrale Fragestellung

Die vorliegende Arbeit gründet auf diesem konkreten Anwendungsbezug. Es handelt es sich um eine Evaluationsstudie zur automatischen Indexierung mit der Software MindServer der Firma Recommind in der sozialwissenschaftlichen Fachinformation. Den Kern der Arbeit bilden vier Testläufe mit dem MindServer auf der Materialgrundlage der Literaturdatenbank SOLIS. Unterschiedliche Systemeinstellungen und Trainingskorpora wurden auf ihren Einfluss auf das Indexierungsergebnis hin untersucht. Dabei wurde in Anlehnung an das Schalenmodell (siehe oben) ein Vergleich der Indexierungsergebnisse zwischen Kern- und Randbereichen der Datenbank SOLIS vorgenommen. Auf der Grundlage der eingesetzten Software wurde davon ausgegangen, durch den Aufbau fachteilgebietsspezifischer Kontext- und Konzepträume das Indexierungsverhalten der Software signifikant verbessern zu können. Zur Evaluation, die sowohl auf der Ebene einer ausgewählten Dokumentenkollektion als auch auf der Ebene der Einzeldokumente erfolgte, wurden hierzu intellektuell generierte Vergleichsdaten herangezogen.

Hintergrund der Arbeit bildet das Ziel, mit der Indexierungssoftware MindServer, eine Vor-Indexierung bei der inhaltlichen Erschließung insbesondere für die Randbereiche der Datenbank einzuführen. Diese lassen sich jedoch nur bedingt bestimmten Datenzulieferern zuordnen (siehe hierzu Kap. 3.1). Nachdem bereits eine Verbindung zwischen dem Produktionssystem a-DIS │BMS und der Indexierungssoftware geschaffen wurde, sind damit die technischen Voraussetzungen bereits erfüllt. Nach Beendigung der Formalerschließung lassen sich Titel und Abstract, umgewandelt in ein XML-Format, von der Indexierungssoftware automatisch erschließen und in das Produktionssystem zurückspielen. Manuell lassen sich damit die Vorschläge für Klassifikation und Deskriptoren vom Indexierer entweder übernehmen oder verwerfen. Die Indexierungsvorschläge bleiben gleichwohl für nachfolgende Auswertungen erhalten.

Bezogen auf das konkrete Vorgehen wurde zu Beginn ein Testlauf durchgeführt, bei dem die Standardeinstellungen der Software übernommen wurden. In einem zweiten Schritt wurden diese Standardeinstellungen geringfügig

modifiziert. Unter anderem wurden *cut-off* Levels für die Vergabe der Deskriptoren und der Klassifikationsnotationen eingeführt sowie das Verhältnis zwischen *Recall* und *Precision* (siehe hierzu weiter unten) zugunsten der *Precision* verändert. In einem nächsten Schritt wurden fachteilgebietsspezifische Versionen des MindServer für die Kernbereiche sowie ein Randgebiet der Datenbank aufgebaut. Diese wurden erneut sowohl unter Verwendung der Standardeinstellungen als auch unter Einsatz modifizierter Systemeinstellungen getestet.

Aufbau der Arbeit

Vor diesem Hintergrund ist die vorliegende Arbeit in vier Teile gegliedert. Der erste Teil bildet eine eher allgemein gehaltene Annäherung an die Bereiche Wissensrepräsentation und Information Retrieval. Im Mittelpunkt stehen Ablauf und Problembereiche des Indexierungsvorgangs. Im zweiten Teil wird auf automatische Indexierungsverfahren eingegangen. Im Anschluss an Ausführungen zu den unterschiedlichen Verfahrensansätzen werden zentrale Standardmaße zu ihrer Evaluation vorgestellt. Teil III behandelt den konkreten Anwendungskontext der Arbeit. Im Anschluss an einen Überblick über die Erschließungsinstrumente, wie sie bei GESIS │Leibniz-Institut für Sozialwissenschaften zum Einsatz kommen, wird die Verfahrensweise der Indexierungssoftware MindServer vorgestellt. Hieran folgen Darstellung und Auswertung der durchgeführten Testläufe. Im abschließenden vierten Teil werden die Untersuchungsergebnisse im Hinblick auf weiteren Forschungsbedarf und die praktische Umsetzung diskutiert sowie das Vorgehen bei der Evaluation kritisch reflektiert.

Teil I: Grundlagen und Problemfelder der inhaltlichen Erschließung

Im Folgenden wird in den allgemeinen Kontext der Arbeit eingeführt. Es werden zentrale Begrifflichkeiten und Instrumente der Wissensrepräsentation vorgestellt und ihr Einsatz im Rahmen der Inhaltserschließung und des Information Retrieval problematisiert.

1.1 Wissensrepräsentation – Information Retrieval

Der Begriff Wissensrepräsentation steht für Methoden und Techniken, Wissen durch die Anwendung von Ordnungssystemen abzubilden, um es in strukturierter Form für die Suche zugänglich zu machen. Repräsentation von Wissen ist somit unmittelbar mit dem Suchen und Finden von Information, dem Information Retrieval, verbunden. Dabei reicht die Entwicklung von Verfahren zum Aufbau von Informationen *über* Dokumente zur Erleichterung der Informationssuche – etwa in Form von Aufstellungssystematiken – bis in die Antike zurück. Neuere Verfahren der Wissensrepräsentation bilden digitale Datenbanken sowie Folksonomies als eine Form der freien Schlagwortvergabe durch die Nutzer. Im Kern geht es dabei jeweils darum, für eine dokumentarische Einheit ein Surrogat zu erstellen, das stellvertretend für das Dokument steht. Die dokumentarische Bezugseinheit wird in eine Dokumentationseinheit, die die zentralen Metadaten des jeweiligen Dokuments enthält, umgewandelt.

Das zentrale Verfahren der Wissensrepräsentation bildet das Indexieren oder Erschließen. Inhalte, die in einem Informationsobjekt vorliegen, werden in eine Indexierungssprache (z.B. Deskriptoren eines Thesaurus oder Notationen einer Klassifikation) übersetzt, um sie innerhalb eines Suchsystems recherchierbar und damit (wieder)auffindbar zu machen. Die Beziehung zwischen Suchanfrage und Antwortverhalten eines Suchsystems wird anhand von Information-Retrieval-Modellen konzeptualisiert. Sie „spezifizieren, wie zu einer Anfrage die Antwortdokumente aus einer Dokumentensammlung bestimmt werden“ (Fuhr 2004[5]: 207).

Eine zentrale Unterscheidung lässt sich zwischen sogenanntem Exact-Match- und Best-Match-Retrieval treffen. Bei Ersterem basiert der Abgleich zwischen Sucheintrag und Ergebnismenge auf einer binären Unterscheidung zwischen Treffern, die ganz genau der Formulierung der Suchanfrage entsprechen, und Nicht-Treffern. Sie bilden die gesamte restliche Dokumentenmenge. Demgegenüber steht das Best-Match- oder auch Partial-Match-Retrieval für eine Ergebnisanzeige, die die Ähnlichkeit der Treffer zur Suchanfrage entsprechend eines sogenannten Relevance Ranking abbildet. Ein solches Verfahren bei der Ergebnisanzeige kann den Nutzer entlasten, indem er durch eine gut aufbereitete Anzeige des Trefferumfelds möglicherweise für ihn relevantere Dokumente findet, ohne dafür die exakte Suchanfrage beherrschen zu müssen.[9] Vagheit und Unschärfe der Suchanfragen (siehe hierzu weiter unten) lassen sich somit besser abbilden (vgl. Bertram 2005: 84).[10]

1.2 Ablauf und Instrumente der inhaltlichen Erschließung

Der Indexierungsvorgang steht für eine Form der inhaltlichen Erschließung über Repräsentation. Für ein Dokument wird eine Repräsentation erstellt, die den Inhalt beschreibt und dem Dokument zugeordnet wird (vgl. Nohr 2004[5]). Herkömmlicherweise gründet dieses Vorgehen auf einem zweistufigen Prozess aus Inhaltsanalyse und Inhaltsdarstellung. In der ersten Phase muss der Inhalt des Dokuments zunächst intellektuell verarbeitet und verstanden werden. Dafür wird entsprechendes Kontextwissen vorausgesetzt. Im Anschluss daran wird der Inhalt dann in eine Indexierungssprache übersetzt. Im Kontext der vorliegenden Arbeit gehören zu dieser Übersetzungsarbeit sowohl die verbale Inhaltserschließung über die Vergabe von Bezeichnungen bzw. Deskriptoren als auch die klassifikatorische Inhaltserschließung durch die Zuordnung einzelner Klassifikationsnotationen. Die Beschreibungsmerkmale eines Dokuments werden somit nicht dem Dokument selbst sondern den oben

[9] Daneben lassen sich Funktionen zur automatischen Empfehlung weiterer Suchterme einsetzen, die bei der Suche behilflich sind, indem sie die Kontextabhängigkeit bzw. Spezifität der Wissensdomäne, wie sie für Thesauri typisch sind, abbilden (Bertram 2005: 225). Unter Umständen lassen sich hierbei auch *Scope Notes* berücksichtigen, in denen etwa Vorläufer-Deskriptoren aufgeführt werden.

[10] Intellektuelles Indexieren steht traditionell im Zusammenhang mit Exact-Match-Retrievalverfahren. Automatische Retrievalverfahren führen im Gegensatz dazu in Form des sogenannten Best-Match-Verfahrens weniger oder besser passende Dokumente auf.

angeführten Dokumentationssprachen (Thesaurus und Klassifikation) entnommen.[11]

Begriffe und Bezeichnungen

Im Zentrum des Erschließungsvorgangs stehen *Begriffe* und *Bezeichnungen*. Begriffe bilden gedankliche Vorstellungen eines Gegenstandes. Sie sind Ergebnis einer Abstraktionsleistung und unabhängig von einer spezifischen sprachlichen Form. Entsprechend ihres Abstraktionsniveaus lassen sich Begriffe in *Kategorien, Allgemeinbegriffe* und *Individualbegriffe* unterteilen. Erstere stehen für Begriffe auf einem sehr hohen Abstraktionsniveau. Dabei hängt es vom Kontext bzw. der jeweiligen Wissensdomäne ab, inwieweit ein Begriff eine Kategorie bildet. Kategorien kommt eine hohe Bedeutung bei der inhaltlichen Erschließung zu, da sie dabei helfen, das Dokumentationsvokabular zu ordnen und zu strukturieren. Allgemeinbegriffe sind im Unterschied dazu auf der nächst tiefer liegenden Abstraktionsstufe angesiedelt. Sie beschreiben eine „Klasse von miteinander verwandten Gegenständen, die wesentliche Merkmale gemein haben“ (vgl. Bertram 2005: 33). Individualbegriffe bilden hingegen einzelne Einheiten, die sich nicht weiter sinnvoll spezifizieren lassen.[12]

In Bezug auf ihre Komplexität lässt sich zwischen *Einzelbegriffen* und *Begriffskombinationen* unterscheiden. Während erste aus einem einzelnen Begriff bestehen (z.B. Migration), weisen Begriffskombinationen mehrere begriffliche Einheiten auf (z.B. Migrationspolitik), die im Deutschen etwa durch Komposita oder auch durch Mehrwortbenennungen gebildet werden.

Begriffe können jeweils entsprechend ihrer Merkmale in unterschiedlicher Beziehung zueinander stehen. Hier lässt sich zwischen *Äquivalenz- Hierarchie-* und *Assoziationsrelation* unterscheiden. Während die Äquivalenzbeziehung für eine Beziehung zwischen zwei bedeutungsgleichen oder gleichgesetzten Begriffen steht, bezeichnet Hierarchierelation ein hierarchisches Ver-

[11] In diesem Zusammenhang wird von einem Additions- im Gegensatz zu einem Extraktionsverfahren gesprochen.
[12] Jutta Bertram (2005: 34) bringt zur Veranschaulichung der Unterscheidung nach Abstraktionsstufen das Beispiel Institution (Kategorie) – Partei (Allgemeinbegriff) – SPD (Individualbegriff).

hältnis aus einem über- und einem untergeordneten Begriff. Für das Retrieval ist hierüber eine Ausdehnung oder eine Einschränkung der Ergebnismenge möglich. Letztere Beziehungen gründen schließlich auf einer in irgendeiner Form vorhandenen inhaltlich-thematischen Assoziation der Begriffe zueinander (vgl. ebd.).

Schließlich lassen sich zwei weitere Formen von Beziehungen der Begriffe zueinander unterscheiden. Gemeint ist die Unterscheidung zwischen *paradigmatischen* und *syntagmatischen* Begriffsbeziehungen. Erstere stehen für Begriffsbeziehungen, die unabhängig vom konkreten Dokument auf der Grundlage eines Ordnungs- oder Sprachsystems hergestellt werden. Darunter fallen die obigen drei unterschiedlichen Formen von Begriffsbeziehungen. Letztere stehen für Beziehungen, die sich aus der konkreten Sprachanwendung in einem Dokument bzw. dem gesamten Dokumentbestand vor dem Hintergrund der begrifflichen Ansetzungsformen innerhalb der eigenen Dokumentationssprache ergeben.[13]

Bezeichnungen stehen hingegen für sprachliche Repräsentationen eines Begriffs. Sie können unterschiedliche Formen annehmen. Hier lässt sich zwischen *Benennungen* als natürlichsprachige Ausdrucksformen für einen Begriff und künstlichsprachigen Bezeichnungen, etwa Notationen, differenzieren.

Der Umstand, dass die Beziehung zwischen Begriff und Bezeichnung nicht immer eindeutig ist, birgt besondere Herausforderungen für den Dokumentationsprozess. Zum einen kann ein und dieselbe Bezeichnung verschiedene Begriffe repräsentieren. Diese „Beziehung zwischen übereinstimmenden Benennungen für unterschiedliche Begriffe“ (DIN 2342/1: 1992,3 nach Stock/Stock 2008: 54) wird als *Homonymie* bzw. Bedeutungsvielfalt, bezeichnet. Bei der Recherche kann dies Ballast erzeugen und somit auf Kosten der Präzision des Suchergebnisses gehen. Erst die Kombination mit anderen Suchtermen kann unter Umständen zu einer sinnvollen Einschränkung des

[13] Ein Beispiel hierfür sind die Begriffe *Biographie* und *Lebenslauf*. Während im Thesaurus Sozialwissenschaften Biographie als Unterbegriff zu Lebenslauf angesetzt wird (paradigmatische Begriffsbeziehung), ist denkbar, dass beide Begriffe im Dokumentbestand synonym verwendet werden (syntagmatische Begriffsbeziehung). In diesem Fall ließe sich über eine Überarbeitung des Thesaurus nachdenken.

Suchraums führen. Zum anderen können unterschiedliche Bezeichnungen ein und denselben Begriff repräsentieren. Für diese „Beziehung zwischen Benennungen, die denselben Begriff bezeichnen“ (ebd.), steht der Ausdruck *Synonymie*, bzw. Benennungsvielfalt. Ihr lässt sich etwa mit Hilfe von Vorzugsbenennungen, die für den Nutzer transparent sein müssen, begegnen.[14] Hinzu kommt die Beziehung zwischen Benennungen, die sich in ihrer Bedeutung überlappen. Diese Beziehung kann bei der Recherche Verlust erzeugen, da unter Umständen nicht sämtliche relevanten Dokumente angezeigt werden. Auch hier kann ein Vorschlagsdienst zur Suchraum-Erweiterung dieses Problem eingrenzen.[15]

Zusätzlich zur Uneindeutigkeit in der Beziehung zwischen Begriff und Bezeichnung treten im Dokumentationsprozess Übersetzungsprobleme zwischen Urheber, Informationsvermittler und Nutzer hervor. So sind im Indexierungs- wie auch im Retrieval-Prozess mehrere Übersetzungsschritte enthalten, die äußerst voraussetzungsvoll und fehleranfällig sind.

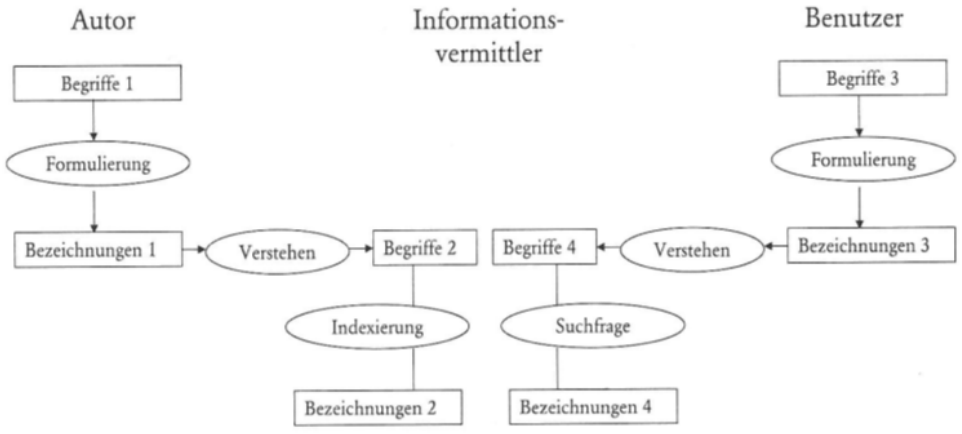


Abbildung 1: Sprachliche Transformationsprozesse im Dokumentationsprozess (Wersig 1978)

Zu Beginn übersetzt der Autor Begriffe (1) in Bezeichnungen (1). Der Informationsvermittler versucht diese zu verstehen. Er übersetzt sie seinerseits in Begriffe (2) und ordnet diesen beim Indexieren Bezeichnungen (2) zu. Die

[14] Sind Vorzugsbenennungen bzw. nicht zugelassene Deskriptoren nicht transparent, kann daraus für den Nutzer eine Terminologiebarriere entstehen. Durch den spezifischen Fachjargon können bestimmte Sucheinstiege ins Leere laufen.

[15] Weitere Problemfelder für den dokumentarischen Prozess im Zusammenhang mit natürlicher Sprache bilden u. a. die Vielfalt an Schreibweisen und Wortformen. Hierauf wird im Kontext automatischer Erschließungsverfahren weiter unten eingegangen.

Inhalte eines Dokuments müssen somit zunächst verstanden werden. Hierfür ist ein entsprechendes Kontext- bzw. Vorwissen des Indexierers unabdingbar. Daneben ist es Aufgabe des Indexierers, sich zu fragen, für welchen Nutzer das Dokument relevant sein könnte, und in welcher Form der spezifische Nutzerkreis eine Suchanfrage formulieren könnte, um das vor sich liegende Dokument zu finden. Je homogener die angenommene Nutzergruppe erscheint, umso eindeutiger kann die Suchanfrage (re)konstruiert werden. Weitere Übersetzungsschritte erfolgen auf Seiten des Nutzers. Er formuliert die Begriffe (3), zu denen er recherchiert, in eine Suchanfrage um und wählt dafür Bezeichnungen (3) aus.[16] Dieses Vorgehen muss der Informationsvermittler ebenso berücksichtigen.[17] Somit muss er das Rechercheanliegen des Nutzers verstehen und in Begriffe (4) umsetzen und die Begriffe schließlich in Bezeichnungen (4) übersetzen (vgl. Bertram 2005).

Klassifikation, Thesaurus und Abstract

Klassifikation, Thesaurus und Abstract bilden die zentralen Instrumente der Inhaltserschließung. Erste beiden stehen ihrerseits für verschiedene Dokumentationssprachen. Handelt es sich bei einem Thesaurus um ein natürlichsprachiges Ordnungssystem, das eine feine inhaltliche Erschließung ermöglicht, stellt eine *Klassifikation* ein künstlichsprachiges Hilfsmittel zur Groberschließung dar, das einzelne Fachgebiete oder breiter gefasste Gegenstandsbereiche systematisch ordnet. Zentrales Prinzip einer Klassifikation ist die Bildung von Klassen, die durch Notationen repräsentiert werden, und sich

[16] In diesem Zusammenhang lässt sich zwischen einem konkreten und einem problemorientierten Informationsbedarf unterscheiden. Während ersterer auf eine Fakteninformation abzielt und die thematischen Grenzen klar abgesteckt sowie die Formulierung der Suchanfrage durch exakte Terme möglich ist, zeichnet sich ein problemorientierter Informationsbedarf vornehmlich dadurch einen iterativen Suchprozess aus (vgl. Gödert/Lepsky 1997). „Das grundlegende Problem ist dabei, etwas klar zu denken und zu formulieren, was man nicht oder zumindest nicht genau kennt“ (Stock 2007: 52). Des Weiteren wird mit dem Zugang zu relevanten Dokumenten mitunter ein neuer, modifizierter Informationsbedarf geweckt. Die Literaturrecherche bildet hier einen iterativen Prozess und kann ihrerseits selbst zur Wissenserweiterung beitragen.

[17] Die subjektive Konturierung von Suchanfragen bringt auch die terminologische Unterscheidung zwischen Informationsbedarfen und Informationsbedürfnissen zum Ausdruck. Während erstere dafür stehen, dass aus einer (relativ) objektiven Betrachtungsperspektive ein Sachverhalt durch eine bestimmte Information „gelöst“ wird, steht der Begriff des Informationsbedürfnisses dafür, dass eine Einschätzung, inwiefern für den konkreten Nutzer die Information tatsächlich eine befriedigende Antwort auf seine Suchanfrage darstellt, nur unter Berücksichtigung seines jeweiligen Vorwissens etc. getroffen werden kann.

inhaltlich nach Möglichkeit deutlich voneinander sowie nach verschiedenen Hierarchiestufen (Haupt- und Unterklasse) unterscheiden lassen.

Im Gegensatz dazu zeichnet sich ein *Thesaurus* durch die Ordnung der Beziehungen zwischen Begriffen und Bezeichnungen aus. Dies schafft die Voraussetzung dafür, durch die Vergabe der Bezeichnungen den Inhalt eines Dokuments möglichst eindeutig wiederzugeben und das Dokument auffindbar zu machen. Zentrales Prinzip bei der Erstellung eines Thesaurus ist daher die Disambiguierung. Dazu gehört, sowohl sämtliche Synonyme einer Bezeichnung zu erfassen als auch Homonyme bzw. Polyseme besonders zu kennzeichnen. Für jeden Begriff wird dadurch lediglich eine Bezeichnung, ein sogenannte Deskriptor, als Teil des Gebrauchsvokabulars vergeben, der diesen eindeutig repräsentieren soll.[18] Deskriptor und zugehörige Synonyme bilden zusammengenommen eine Äquivalenzklasse. Neben diesen Beziehungen zwischen bedeutungsgleichen bzw. -gleichgesetzten Bezeichnungen weist ein Thesaurus in der Regel weitere Relationen zwischen den aufgeführten Begriffen und Bezeichnungen auf, wodurch ein weit verzweigtes Beziehungsgeflecht zwischen den Begriffen und Bezeichnungen entsteht (siehe hierzu ausführlich Tab. 8 im Anhang).

Zentrales Prinzip bei der Anwendung eines Thesaurus bildet hingegen die *Begriffsgleichordnung*. Sie steht dafür, dass jeder Deskriptor mit jedem anderen kombiniert werden kann. Durch die Aneinanderreihung von Deskriptoren werden die unterschiedlichen Dimensionen eines Dokuments darstellbar. Für diese Anwendungsmöglichkeit steht auch der Begriff der *Postkoordination*: komplexe Sachverhalte werden zerlegt und lassen sich durch die Aneinanderreihung von Deskriptoren wiedergeben. In diesem Zusammenhang stellt sich somit erneut bereits beim Aufbau eines Thesaurus die Frage, inwieweit komplexe Bezeichnungen, im Deutschen handelt es sich hierbei zumeist um Komposita, prä- oder postkombiniert in den Thesaurus aufgenommen werden sollen. Ob also der Nutzer die Begriffe bereits zusammengesetzt vorfindet, oder aber beim Suchen selbst zusammensetzen muss. Hiervon hängen we-

[18] Zusätzlich zum Gebrauchsvokabular eines Thesaurus existiert das Zugangsvokabular. Es setzt sich aus den sogenannten Nicht-Deskriptoren zusammen und soll zum Gebrauchsvokabular hinführen. Neuartige Bezeichnungen, die sich in der Domäne des Thesaurus erst noch etablieren, stellen das sogenannte Kandidatenvokabular dar.

sentlich der Umfang eines Thesaurus und die Vielfalt an Beziehungen, die unter den Bezeichnungen möglich sind, ab. Es gilt, sich hierbei bewusst zu machen, dass eine Zerlegung der Bezeichnungen auf der einen Seite zusätzliche Einstiegsmöglichkeiten bietet.[19] Auf der anderen Seite geht sie allerdings auf Kosten der Spezifität und Genauigkeit bzw. Präzision bei der Recherche. Gleichzeitig erhöht sich die Vergabehäufigkeit der Deskriptoren, wodurch Fehlverknüpfungen und damit Ballast bei der Recherche entstehen können (vgl. Bertram 2005).

Das *Abstract* stellt hingegen eine verdichtete Wiedergabe des Dokumentinhalts in seinem Kontext dar. Wie Klassifikation und Thesaurus unterstützt es somit den Nutzer bei der Entscheidung, ob eine Publikation für ihn von Relevanz ist. In Fällen, in denen Abstracts über die Freitextsuche durchsuchbar sind, erfüllen sie zusätzlich eine Zugangsfunktion, die ansonsten in der Regel den Indextermen bzw. der Klassifikation zukommt. Als allgemeine Richtschnur zur Erstellung eines Abstracts gelten bestimmte Vorgaben, die untereinander zum Teil in einem Spannungsverhältnis stehen. So sollen Abstracts generell vollständig aber gleichzeitig kurz gehalten sowie verständlich und neutral geschrieben sein. Nach dem Abstraktionsniveau lässt sich zwischen *indikativem* Abstract, das den Hauptgegenstand einer Publikation und das Untersuchungsvorgehen wiedergibt, und *informativem* Abstract, das zusätzlich auf das Erkenntnisinteresse sowie Hypothesen und Ergebnisse eingeht, unterscheiden (vgl. Kuhlen 2004).[20]

1.3 Problemfelder der inhaltlichen Erschließung

Wie sich bereits in den vorangegangenen Ausführungen andeutet, werden sowohl Indexierung als auch Retrieval von zahlreichen Faktoren beeinflusst. Dazu zählen nicht nur die jeweils konkrete dokumentarische Bezugseinheit, die spezifische Wissenschaftsdisziplin und die dazugehörige Dokumentati-

[19] In diesem Zusammenhang wird auch von einem gleichordnenden im Gegensatz zum syntaktischen Indexieren gesprochen. Es zeichnet sich durch die Indexiermethode der Postkoordination aus, bei der die Sachverhalte in ihre begrifflichen Komponenten zerlegt werden: z.B. Migration AND Theorie anstellte von Migrationstheorie. In der Praxis kommt es wesentlich häufiger zum Einsatz, da es zeitlich und ökonomisch weniger aufwendig sowie in Bezug auf das nötige Vor.- bzw. Kontextwissen weniger voraussetzungsvoll ist.

[20] Auf das Abstracting als Erschließungsmethode wird weniger ausführlich eingegangen, da im konkreten Anwendungsfall, auf dem diese Arbeit beruht, keine automatische Zusammenfassung von Textdokumenten im Sinne eines automatischen Abstracting vorgenommen wird.

onssprache sowie die Datenbank, in die das Indexat einfließt, sondern auch das spezifische Recherchebedürfnis des Datenbanknutzers, das konkrete Indexierungsvorgehen und nicht zuletzt der Indexierer selbst. So handelt es sich bei der inhaltlichen Erschließung um einen deutlich subjektiv geprägten Vorgang. Obgleich beim Aufbau der Indexierungssprachen versucht wird, ähnlich gelagerte Sachverhalte auch in einer ähnlichen Weise wiederzugeben, um sie dadurch (wieder)auffindbar zu machen, schließt die intellektuelle Inhaltserschließung, wie jeder Prozess, der an ein Verstehen geknüpft ist, ein subjektives, interpretatives Moment mit ein.[21]

In diesen Bereich fällt auch die Problematik der Indexierungskonsistenz. Sie bezeichnet das Maß an Übereinstimmung zwischen unterschiedlichen Indexaten derselben Vorlage (vgl. Stock/Stock 2008) und betrifft zum einen die Konsistenz zwischen den Indexaten eines Dokuments, das von unterschiedlichen Personen erschlossen wurde, die sogenannte Inter-Indexiererkonsistenz. Zum anderen verweist sie auf die Indexierungskonsistenz ein und derselben Person bei der Erschließung eines Dokuments zu unterschiedlichen Zeitpunkten, die sogenannte Intra-Indexiererkonsistenz. Zusätzlich lässt sich die Indexierungskonsistenz danach untersuchen, inwieweit gleiche bzw. ähnliche Inhalte in verschiedenen Dokumenten in gleicher Weise indexiert werden. Dieses Maß an Konsistenz, das explizit das Retrieval betrifft, wird als Indexer-Anfrage- bzw. Indexer-Nutzer-Konsistenz bezeichnet (vgl. ebd.).[22]

Die Konsistenz zwischen unterschiedlichen Indexierern wurde wiederholt in Studien untersucht. Dabei zeigte sich, dass sich Erschließungsergebnisse von erfahrenen Indexierern tendenziell einander annähern, während die Konsistenz von Indexierungen, die von Laien vorgenommen werden, sehr viel stärker variiert (vgl. David/Giroux 1995 sowie Saarti 2002 nach Groß 2010). Gleichwohl werden auch für die Inter-Indexiererkonsistenz von erfahrenen Sacherschließern unterschiedlicher Fachgebiete Konsistenzwerte von (lediglich) 45 bis knapp 80 Prozent nachgewiesen (vgl. ebd.). Besonderen Einfluss

[21] Gleiches gilt im Übrigen auch für die Evaluation (siehe hierzu weiter unten).
[22] Die Berechnung der Indexierungskonsistenz stellt auch für die Evaluation automatischer Erschließungsverfahren eine gängige Maßzahl dar. Hierbei wird das Maß an Übereinstimmung zwischen intellektuell sowie automatisch generiertem Indexat betrachtet (siehe hierzu Kap. 2.3).

auf die Indexierungskonsistenz haben diesen und weiteren Untersuchungen zufolge vor allem die Unterscheidung zwischen Haupt- und Randthemen, die Thesaurusstruktur sowie allgemein das Fachgebiet. So sei die Indexierungskonsistenz in den Geistes- und Sozialwissenschaften besonders gering (vgl. Xu 2008 sowie Saarti 2002 nach Groß 2010).[23]

Daneben, so klang bereits an, sind sowohl Indexierung als auch Retrieval davon geprägt, dass die Repräsentation von Inhalten – bei der Auswahl inhaltskennzeichnender Terme im Zuge der Erschließung sowie bei der Eingabe einer Suchanfrage – stets von der Unschärfe und Vagheit der natürlichen Sprache begleitet ist. Diese Uneindeutigkeit wird dadurch gesteigert, dass zumeist die Beziehungen zwischen den ausgewählten Bezeichnungen der Repräsentation nicht deutlich werden. So bedeutet jede Form von Repräsentation auch einen Informationsverlust.[24] Techniken und Verfahren der Repräsentation von Wissen sind somit an eine Interpretation des vorliegenden Wissens bzw. der vorliegenden Informationen geknüpft. Einzig bei der „reinen" Extraktion von Dokumentinhalten ließe sich behaupten, dass die Notwendigkeit einer eigenen Interpretation des Dokumentinhalts weniger erforderlich ist. Doch selbst hierbei unterliegt die Auswahl von Textpassagen einer Interpretation.

[23] Die Werte, die nach unterschiedlichen Formeln (Hooper/Rolling) berechnet wurden (siehe hierzu weiter unten), liegen bei Xu (2008) zwischen 60 und 70 Prozent.

[24] Berücksichtigt werden diese Überlegungen in dem Konzept der kognitiven Modelle (vgl. Stock 2007). Demnach sind nicht nur die Suchanfragen der Nutzer von den jeweiligen Wissensbeständen sowie Wahrnehmungen und Orientierungen, sondern auch die Instrumente der Wissensrepräsentation von den Vorkenntnissen und Perspektiven ihrer Konstrukteure und Autoren beeinflusst. Die Vorstellung kognitiver Modelle wird dabei auch auf die Verfahren der automatischen Indexierung ausgeweitet. In diesem Zusammenhang stünden kognitive Modelle für die von den Information-Retrieval-Modellen verwendeten Algorithmen.

Teil II: Verfahrensansätze und Evaluationsformen der automatischen Indexierung

Im Gegensatz zur intellektuellen oder auch manuellen Indexierung, wie sie im vorangegangenen Teil behandelt wurde, steht das automatische Indexieren für eine Inhaltsanalyse und -darstellung auf maschinellem Wege. In gleicher Weise wie beim intellektuellen Vorgehen lässt sich in Bezug auf die Ermittlung der Indexterme zwischen Extraktions- und Additionsverfahren unterscheiden. Die Indexierungsarbeit wird an ein automatisches Verfahren delegiert. Nicht selten schließt sich daran im Rahmen semiautomatischer Erschließungsverfahren erneut eine intellektuelle Analyse und Nachbearbeitung des automatisch generierten Indexierungsergebnisses an.

Auch wenn intellektuelle und maschinelle Verfahren der Inhaltserschließung miteinander verzahnt sein können, gründen sie auf ganz unterschiedlichen Vorgehensweisen. Zum einen zielt der intellektuelle Ansatz herkömmlicherweise in erster Linie auf die konsistente Wiedergabe des Dokumentinhalts ab, wofür seine Erschließung auf der Bedeutungsebene notwendig ist. Maschinelle Verfahren bewegen sich im Gegensatz dazu alleine auf der sprachlichen Oberfläche der Dokumente. Zum anderen steht bei der intellektuellen Erschließung vor allem die korrekte Repräsentation des Dokumentinhalts im Vordergrund, während automatische Verfahren sehr viel stärker auf die Wiederauffindbarkeit des Dokuments abzielen (vgl. Nohr 2003). Automatische Verfahren zeichnen sich somit in der Regel durch ein Best-Match-Retrieval aus (siehe oben).

Die Verlagerung von einem Exact- hin zu einem Best-Match-Retrieval spiegelt sich auch in der Diskussion um die Entwicklung und Einführung automatisierter Erschließungsverfahren wider. Ausgehend von der Vielfalt sprachlicher Ausdrucksweisen führen Kritiker an, dass es unmöglich sei, diese adäquat über maschinengestützte Verfahren abzubilden. Das zentrale Argument hinter dieser Position lautet, dass es für die Inhaltserschließung unerlässlich sei, bis auf die Bedeutungsebene sprachlicher Zeichen vorzudringen. Erst unter Einbezug des Kontextes, der über ein maschinelles Verfahren nur be-

dingt zugänglich sei, könne die inhaltliche Wiedergabe möglich sein. Dem gegenüber argumentieren Anhänger automatischer Verfahren damit, dass unterschiedliche Untersuchungen durchaus Belege für eine hohe Indexierungsqualität der Verfahren im Sinne der Wiederauffindbarkeit der Dokumente erbracht hätten (vgl. ebd.).

2.1 Zentrale Verfahrensansätze

Verfahren der automatischen Indexierung, die explizit Metadaten produzieren, lassen sich in unterschiedliche Ansätze unterteilen. In Bezug auf die Ermittlung inhaltskennzeichnender Indexterme lässt sich zwischen statistischen und computerlinguistischen Verfahren unterscheiden. Für die sich daran anschließende Zuordnung von Indextermen aus einem kontrollierten Vokabular, wie sie bei einem Additionsverfahren vorgenommen wird, stehen begriffsorientierte Verfahrensansätze. Vielfach werden die genannten Verfahren, die im Folgenden näher ausgeführt werden, miteinander kombiniert.

2.1.1 Statistische Verfahren

Statistische Verfahren stellen die ersten und daher am weitesten entwickelten Verfahrensansätze dar. Sie gründen auf der Annahme, dass die Frequenz, in der ein Term in einem Dokument vorkommt, etwas über die Bedeutung dieses Terms aussagt. Diese Hypothese stellte als erster Ende der 1950er Jahre der deutsche Informatiker Hans Peter Luhn auf. Dahinter steckte die Vorstellung, die bedeutungstragenden Inhalte eines Textes über statistische Verfahren ermitteln zu können. Luhn knüpfte hierbei auf Überlegungen des US-amerikanischen Linguisten George Kingsley Zipf an, der die Beziehung zwischen der Häufigkeit eines Wortes und seiner Frequenz in einem Text untersuchte und hierfür eine Gesetzmäßigkeit, das sogenannte Zipfsche Gesetz, formulierte.

Daneben liegen den statistischen Verfahrensansätzen weitere Annahmen zugrunde. Zum einen wird davon ausgegangen, dass aus den vorhandenen Termen eines Dokuments eine bestimmte Selektion getroffen werden muss, da sich nicht alle als Indexierungsterme eignen. Zum anderen wird die Auffassung vertreten, dass die in dieser Form ausgewählten Indexierungsterme in

unterschiedlicher Weise zu der Bedeutung eines Textes beitragen. Die Terme werden daher unterschiedlich gewichtet, was die zentrale Voraussetzung für die Anwendung eines Best-Match-Retrievalverfahrens bildet. Zentrales Instrument zur Ermittlung der Bedeutungsrelevanz ist die Häufigkeit, mit der ein Term im Dokument vorkommt.

Aus diesen Überlegungen werden weitere Annahmen abgeleitet. Bezogen auf ein einzelnes Dokument wird davon ausgegangen, dass eine hohe Termfrequenz auch mit einer hohen Bedeutungsrelevanz korreliert. Unter Berücksichtigung der gesamten Dokumentensammlung wird angenommen, dass weniger häufig vorkommende Wörter sich eher als Indexierungsterme eignen – sie weisen einen höheren sogenannten Diskriminanzeffekt auf (vgl. Nohr 2003). Beide Annahmen fließen in die Ermittlung der inversen Dokumenthäufigkeit (IDF) ein, wie sie auf der Grundlage von Untersuchungen zur Analyse von Worthäufigkeiten in großen Textmengen von Karen Sparck Jones (1973 nach Stock 2007) eingebracht wurde. Hiernach verhält sich die Bedeutsamkeit eines Terms proportional zur Häufigkeit seines Auftretens in einem einzelnen Dokument, hingegen umgekehrt proportional zur Gesamtanzahl jener Dokumente, in denen dieser Term auftritt (vgl. Nohr 2004).

In diesem Zusammenhang lässt sich eine grobe Unterteilung in hoch-, mittel- und niedrigfrequente Terme vornehmen. Während sowohl für Terme mit einer hohen und mit einer niedrigen Frequenz angenommen wird, dass sie nur wenig bedeutungstragend sind, wird vor allem für mittelfrequente Terme von einer hohen Bedeutungsrelevanz ausgegangen. So fallen wenig relevante Terme vermutlich in den Randbereich, während sowohl Terme, die für ein Fachteilgebiet besonders häufig vergeben werden, als auch Terme, die als Grundbausteine der Satzkonstruktion wichtig sind, ausgesprochen häufig vorkommen. Um den mittleren Frequenzbereich für die Indexierung von Dokumenten nutzbar zu machen, werden spezifische Schwellenwerte ermittelt, die es zu über- bzw. unterschreiten gilt. Der Berechnung der entsprechenden Schwellenwerte kommt für die Anwendung statistischer Verfahrensansätze somit eine hohe Bedeutung zu.

Daneben fließen weitere Faktoren in die Gewichtung eines Terms ein. Hierzu zählt zum einen die Position des Terms im Dokument. Terme, die im Titel aufgeführt werden, können höher bewertet werden, als Terme, die im Fließtext eines Dokuments oder Abstracts vorkommen. Zum anderen wirkt sich die adäquate Erkennung eines Terms auf dessen Gewichtung aus. In diesem Zusammenhang ist als Beispiel für die Kombination statistischer mit informationslinguistischen Verfahren etwa die korrekte Zurückführung eines Terms auf seine Grundform von besonderer Relevanz. Dies betrifft sowohl die Zuordnung unterschiedlicher Formen eines Wortes auf einen Term als auch die Zerlegung von Komposita. Und auch die Analyse des gemeinsamen Auftretens bestimmter Wörter ist von Bedeutung. So können bestimmte Wortgruppen oder Kollokationen als zusammengehörig definiert werden.

In der Literatur werden zur Anwendung statistischer Verfahrensansätze bestimmte Bedingungen angeführt. Zum einen sollte eine ausreichende Textgrundlage, etwa in Form von Volltexten oder Abstracts, vorhanden sein. Dabei zeigte sich, dass Kurzreferate, die den Inhalt eines Textes wiedergeben, für die Anwendung statistischer Verfahrensansätze besser geeignet sind als Volltexte (vgl. Nohr 2004). Zum anderen trägt es zu einer höheren Indexierungsgüte automatischer Verfahren bei, wenn es sich bei der Dokumentenkollektion um eine fachlich homogene Textsammlung handelt, in der die jeweiligen Terme in ähnlicher Weise verwendet werden. Schließlich ist eine ausreichend große Datenkollektion für den Einsatz statistischer Verfahren notwendig (vgl. ebd.).

Im Kontext von Information-Retrieval-Systemen werden durch die Verwendung der Termfrequenz, wie sie sich über statistische Verfahren ermitteln lässt, auch Ähnlichkeiten zwischen Dokumenten berechnet. Sie bilden die Grundlage des Vektorraummodells (siehe hierzu das Prinzip der latenten probabilistischen Berechnung von Dokumentähnlichkeiten unter Kap. 3.2). Vektorräume sind vieldimensionale Räume, deren Anzahl von Dimensionen sich aus der Gesamtanzahl der vergebenen Indexierungstermen innerhalb einer Dokumentenkollektion berechnet. Jedes einzelne Dokument erscheint in diesem mehrdimensionalen Raum entsprechend der erfolgten Indexierung als ein Vektor und bekommt somit einen festen Platz in diesem Modell zugewiesen.

Inhaltlich ähnliche Dokumente sind sehr nah zueinander angeordnet. Die Ähnlichkeit der Dokumente errechnet sich dabei aus ähnlich gelagerten Termfrequenzen übereinstimmender Indexierungsterme. Zentral für die Ermittlung der inhaltlichen Ähnlichkeit der Dokumente sind auch hier wieder die mittelfrequenten Indexterme. Auf dieser Grundlage lassen sich auf den Inhalt bezogene Cluster, als „Mengen ähnlicher Dokumente, die bereits bei der Anlegung der Kollektion ermittelt und gespeichert werden“ (ebd.: 219), bilden. Analog erhält auch jede Suchanfrage, die an ein Retrieval-System gestellt wird, einen Punkt innerhalb des Vektorraummodells zugeteilt. Für die Ergebnisanzeige von Dokumenten, die zu der Suchanfrage passen, wird ein Abgleich zwischen Dokument- und Anfragevektor vorgenommen. Diejenigen Dokumente, die innerhalb des Vektorraums der Suchanfrage am nächsten stehen und somit inhaltlich am ähnlichsten sind, werden schließlich als Suchergebnisse angezeigt. Aufgrund ihrer Ähnlichkeit wird auf ihre Relevanz für die Suchanfrage geschlossen.

2.1.2 Linguistische Verfahren

Computer- bzw. informationslinguistische Verfahren ermitteln Indexterme hingegen auf der Basis sprachlicher Gesetzmäßigkeiten. Hierzu beziehen sie unter anderem den Bereich der Morphologie ebenso mit ein wie die Syntax, um die Vielfalt sprachlicher Phänomene zu reduzieren. Voraussetzung dafür ist sowohl eine erfolgreiche Spracherkennung als auch ein sprachlich homogener Dokumentbestand (vgl. Bertram 2005). Eine Reihe von Textanalyseschritten, die von linguistischen Verfahren vorgenommen werden und zum Teil auch in Softwarelösungen integriert sind, welche auf statistischen Verfahren basieren, sollen im Folgenden aufgeführt werden.

Ein erster Arbeitsschritt linguistischer Verfahren sieht zumeist die Eliminierung sogenannter Stopp-Wörter vor. Als solche werden etwa Artikel oder Präpositionen bezeichnet, die keine sinntragende Bedeutung haben. Sie können über hinterlegte Wortlisten erkannt werden, oder statistisch über ihre Häufigkeit in Texten ermittelt werden (vgl. ebd.). Im Durchschnitt machen sie ein Drittel eines Textes aus.

Ebenso bedeutsam ist die Reduktion von Wortformen. Hierzu lassen nicht nur die Rückführung grammatikalischer Flexionsformen auf ihre Grundform (Lemmatisierung) und die Zerlegung unterschiedlicher morphologischer Varianten auf ihre Stammform (Stemming) sondern auch die Zerlegung von Komposita zählen.[25] Generell lässt sich in diesem Zusammenhang zwischen regelbasierten und wörterbuchgestützten Ansätzen unterscheiden, die beide zumeist miteinander kombiniert werden. Regelbasierte Verfahren gründen bei ihrer Analyse auf Regeln, die in Form von Algorithmen aufgestellt werden. Über diese Regeln lassen sich allgemein gehalten Vorschriften formulieren, die gleichzeitig für eine Vielzahl von Anwendungsbeispielen gelten. Dadurch kann der Pflegeaufwand relativ gering gehalten werden. Allerdings können in der Sprache Ausnahmen vorkommen, für die die aufgestellten Regeln nicht greifen. Inwieweit bestimmte Verarbeitungsvorschriften somit einen großen Bereich sprachlicher Phänomene abdecken, hängt somit auch mit der Grammatik und dem Aufbau der jeweiligen Sprache zusammen. Flexionsarme und morphologisch relativ wenig komplexe Sprache, wie etwa das Englische, eignen sich somit für regelbasierte linguistische Verfahren besonders gut – anders als im Fall des Deutschen, das sich durch seine Vielzahl von Kompositabildungen sowie seinen Flexionsreichtum auszeichnet. Hier erscheint mitunter der Einbezug des begrifflichen Umfeldes unabdingbar.[26]

Wörterbuchbasierte Verfahren zeichnen sich hingegen dadurch aus, dass für die linguistische Analyse ein Wörterbuch hinterlegt ist und somit keine den Einzelfall übergreifenden Regeln aufgestellt werden. Diese Form der linguistischen Analyse eignet sich somit etwa für das Deutsche aufgrund der Kompositabildung besser als für das Englische. Gleichwohl zeichnen sich diese

[25] Eine Reduktion auf die Grundform stellt etwa eine Rückführung der Pluralform auf den Singular oder die Rückführung einer bestimmten Deklinationsendung auf den Nominativ dar. Die Reduktion auf die Stammformen steht hingegen für die Rückführung von Verben und Substantiven auf ihre zentralen sinntragenden Einheiten bzw. Silben. Verschiedene Wortformen lassen sich auf diese Weise auf einen einzigen Indexterm reduzieren (vgl. Bertram 2005).

[26] Die Zerlegung von Komposita kann sehr voraussetzungsvoll sein. Bereits einfache Beispiele machen deutlich, wie anspruchsvoll die Zerlegung in sinnvolle Einzelbestandteile des Kompositums ausfallen kann. So ließe sich etwa das Kompositum Glücksautomaten statt in Glücks und Automaten ebenso auch in Glück, Sau sowie Tomaten zerlegen. Daneben kann bei der Dekomposition die gleiche Gewichtung der Einzelbestandteile für die Inhaltsbestimmung irreführend sein. Bertram (2005: 103) bringt als anschauliches Beispiel den Begriff Wissensdurst.

Verfahren durch einen hohen und kontinuierlichen Pflegeaufwand aus, was sich als besonders zeit- und kostenintensiv erweisen kann. Daneben lassen sich nicht für sämtliche Anwendungsfälle vollständige Namenslisten hinterlegen, auf die das System zurückgreifen kann. Hierzu zählen etwa Personen-, Orts- und Institutionennamen.[27]

Ein weiterer Analyseschritt linguistischer Verfahren stellt die Erkennung von Wortgruppen und Phrasen dar. Dies kann sowohl wörterbuchbasiert als auch in Kombination mit einer statistischen Analyse erfolgen. Ziel ist es, Wörter, die aufeinander folgen oder getrennt voneinander in einem Satz stehen und sich inhaltliche aufeinander beziehen, als gemeinsames sinntragendes Element eines Satzes zu erkennen. Anspruchsvoll ist die Erkennung von Wörtern mit einem gemeinsamen Sammelwort sowie von Paraphrasen. In ähnlicher Weise stellen die Erkennung von Wortbindestrichtilgungen sowie die Umformung von Adjektiven in Substantive große Herausforderungen für computerlinguistische Erschließungsverfahren dar.

2.1.3 Begriffsorientierte Verfahren

Während computerlinguistische Verfahren demnach versuchen, geeignete Indexterme auf der Zeichenebene zu ermitteln, dringen begriffsorientierte Verfahren bis auf die semantische Ebene vor. Grundlage hierfür bleibt jedoch – wie bei allen maschinellen Indexierungsverfahren – die sprachliche Oberfläche des Volltextes bzw. Abstracts. Auf der Basis einer Textwortanalyse werden bedeutungstragende Wörter mit einem zugrundeliegenden kontrollierten Vokabular abgeglichen, aus dem die entsprechenden Indexterme abgeleitet werden. Hierfür wird zumeist erneut sowohl auf statistische als auch auf informationslinguistische Verfahren zurückgegriffen. Die Zusammenführung verschiedener Benennungen eines Begriffs, wie sie für die Disambiguierung eine wichtige Voraussetzung darstellt, wird somit erreicht. Dabei ist die Implementierung eines solchen Verfahrens allerdings mit einem hohen Pflegeaufwand verbunden, da das hinterlegte Vokabular an Deskriptoren stets aktu-

[27] Plastisch wird dies etwa an Beispielen von mehrdeutigen Zeichenfolgen. So etwa im Fall von „oder“ (Oder – Fluss) oder „singen“ (Singen – Stadt).

alisiert werden muss. Da dies nur zeitversetzt erfolgen kann, ist die Indexierung tendenziell immer lückenhaft (vgl. Nohr 2004 sowie Siegmüller 2007).

2.2 Formen der Evaluation automatischer Indexierung

Die Bewertung automatischer Erschließungsverfahren erfolgt zumeist im Rahmen von Retrieval-Tests. Gemessen wird, „wie gut die Systeme in der Lage sind, die an sie gestellten Anforderungen zu erfüllen, relevante Dokumente zu liefern und nicht-relevante zurückzuhalten“ (Womser-Hacker 2004: 227). Im Zentrum der Messung steht somit der Retrieval-Output. Die Verfahren der Qualitätsmessung werden ständig weiterentwickelt. Aktuell wird vor allem versucht, die (potentiellen) Nutzer der Systeme stärker in die Bewertung einzubeziehen.

Wie bereits im obigen Zitat anklingt, basiert die Evaluation – angelehnt an das binär strukturierte Boolesche Retrieval – in erster Linie auf einer zweistufigen Bewertungsskala aus relevanten und nicht-relevanten Indexierungsergebnissen. Bezogen auf Ranking-Systeme spielt ebenso in die Ergebnisbewertung hinein, inwieweit die relevantesten Treffer am höchsten gerankt werden. Auf dieser Grundlage werden die beiden Standardmaße zur Messung der Effektivität eines Retrieval-Systems, *Recall* und *Precision* gebildet. Der *Recall* gibt die Vollständigkeit eines Retrievalergebnisses an. Er steht für das Verhältnis zwischen selektierten relevanten Dokumenten und in der Dokumentensammlung insgesamt vorhandenen relevanten Dokumenten.

$$Recall = A / A + B$$

A steht somit für die Anzahl der selektierten relevanten Dokumente. B hingegen umfasst die Menge an relevanten Dokumenten, die nicht in der Treffermenge enthalten sind.

Die *Precision* ermittelt indessen die Genauigkeit eines Retrievalergebnisses. Sie stellt das Verhältnis zwischen selektierten relevanten Dokumenten und der Gesamtanzahl nachgewiesener Dokumente dar.

$$Precision = A / A + C$$

A steht erneut für die Anzahl der selektierten relevanten Dokumente. C umfasst in diesem Fall die Anzahl der insgesamt selektierten Dokumente. Beide Maße ergänzen sich in ihrer Aussage zur Effektivität eines Retrieval-Systems. Gleichwohl stehen sie in einem bestimmten Spannungsverhältnis zueinander. So trifft der *Recall*-Wert einzig eine Aussage dazu, wie vollständig ein Retrievalergebnis ausfällt, ohne die Ballastquote an nicht-relevanten Dokumenten, die ebenso in der Trefferanzeige enthalten sein können, einzubeziehen. Im Gegensatz dazu gibt der *Precision*-Wert allein die Effektivität des Retrieval-Systems an, nicht-relevante Dokumente aus der Trefferanzeige herauszufiltern.

Die zentrale Frage, die bei dieser Form der Bewertung aufgeworfen wird, ist die nach der Definition von Relevanz. Angesichts der Vagheit und Unschärfe bei der Formulierung einer Suchanfrage, wie sie sich insbesondere bei problemorientierten Suchbedarfen zeigen, liegt hierin ein besonderes Kritikpotential gegenüber dieser Form der Evaluation. Es wird „ein Widerspruch zwischen der statistisch-quantitativen Anwendung von Maßen und der relativ unscharfen, nur schwer in quantitativen Kategorien fassbaren Basis der Relevanzbewertung gesehen“ (ebd.: 227f.).[28]

In der vorliegenden Arbeit wird lediglich eine Teilkomponente des Retrieval-Systems untersucht. Im Zentrum steht die *Indexierungs*leistung eines automatischen Erschließungsverfahrens. Betrachtet wird nicht direkt das Retrievalergebnis in Bezug auf den gesamten Bestand einer Dokumentenkollektion, sondern die Vergabe der Deskriptoren sowie die Zuordnung der Klassifikationsnotationen auf der Ebene der einzelnen Dokumente. Erschließungsleistung auf der einen und das Antwortverhalten eines Information-Retrieval-Systems auf der anderen Seite sind gleichwohl auf das Engste miteinander verknüpft. Die Repräsentation des Dokumentinhalts in Form von Deskriptoren und Klassifikation entscheidet über die Auffindbarkeit eines Dokuments und somit über die Effektivität eines Information-Retrieval-Systems.

[28] Während sich bei einer faktenorientieren Suchanfrage relativ leicht beurteilen lässt, ob ein Dokument den Informationsbedarf befriedigen kann oder nicht, ist dies bei einem problemorientierten Verfahren nur bedingt möglich. „Beim problemorientierten Informationsbedarf (…) und den Literaturinformationen als Antwort wird die Beurteilung der Relevanz also durchaus vage ausfallen“ (Stock/Stock 2008: 52).

Bei der Auswertung der automatisch generierten Erschließungsergebnisse wurden die manuellen Indexierungen als Relevanzmaßstab angelegt. Für die nachfolgend beschriebenen Testläufe stellt der *Recall* somit das Verhältnis zwischen (automatisch) selektierten relevanten, d.h. ebenso manuell vergebenen, Deskriptoren bzw. Klassifikationsnotationen und im intellektuell generierten Indexat insgesamt aufgeführten Deskriptoren und Notationen dar. Die *Precision* hingegen steht für den Quotienten aus der Überschneidungsmenge der sowohl intellektuell als auch maschinell vergebenen Deskriptoren bzw. Klassifikationen und der Gesamtanzahl maschinell generierter Deskriptoren bzw. Klassifikationsnotationen (vgl. Groß 2010).

Angesichts gängiger Suchmaschinen, die sich durch große Treffermengen auszeichnen, hat sich die Vorstellung von Qualität in jüngster Zeit verschoben. Eine größere Bedeutung als der Vollständigkeit eines Suchergebnisses – bei umfangreichen Retrievaltests ist bezogen auf den Gesamtdokumentbestand hierfür ohnehin ein Schätzwert erforderlich – kommt der Präzision eines Retrieval-Systems, und damit der Fähigkeit, Ballast herauszufiltern, zu. In der vorliegenden Untersuchung wurde daher bei manchen Testläufen (siehe Teil III) die Präzision des Indexierungsergebnisses – gemäß der manuellen Erschließung – an bestimmten Punkten, den sogenannten *cut-off* Levels, gemessen.[29] Zusätzlich wurde für die Vergabe der Deskriptoren die Berechnung des *R-Precision*-Wertes vorgenommen. Der *cut-off* Level richtete sich hierbei jeweils nach der Gesamtanzahl manuell vergebener Deskriptoren.

Daneben werden weitere Kriterien für die Qualität der Indexierungsergebnisse in die Untersuchung mit aufgenommen. Hierzu zählen die Indexierungsbreite und damit die Anzahl der automatisch generierten Indexierungsvorschläge, sowie die Indexierungskonsistenz, die durch die Indexierungsbreite deutlich beeinflusst wird (vgl. Stubbs et al. 1999). Für die Indexierungskonsistenz wird dabei folgende Formel nach Rolling (1981) verwendet:

$$\text{Indexierungskonsistenz} = 2 * C / A + B$$

[29] Zur Ermittlung der Indexierungsergebnisse für bestimmte *cut-off* Levels ließ sich der Konfidenzwert nutzen, der von dem automatischen Erschließungsverfahren für jeden Deskriptor- und Klassifikationsvorschlag vergeben wird (siehe hierzu weiter unten).

C umfasst in dieser Formel die Menge der übereinstimmenden Indexierungskategorien, während A und B jeweils für die Gesamtanzahl der vergebenen Indexierungskategorien der einzelnen Indexierer steht. Bezogen auf die vorliegende Arbeit umfasst C die Menge an Deskriptoren bzw. Klassifikationsnotationen, die sowohl intellektuell als auch automatisch generiert wurde, während A bzw. B für die Gesamtmenge der intellektuell respektive automatisch vergebenen Deskriptoren bzw. Klassifikationsnotationen steht.[30]

[30] Eine andere Möglichkeit zur Berechnung der Indexierungskonsistenz legte R. S. Hooper (1965) vor. Die Formel lautet: A/B+C−A.

Teil III: Beschreibung und Analyse der durchgeführten Testläufe

Im Folgenden werden die Testläufe, auf denen die vorliegende Arbeit beruht, beschrieben. Einleitend werden die Datengrundlage sowie die Erschließungsinstrumente in der sozialwissenschaftlichen Fachinformation und die Indexierungssoftware vorgestellt. Im Anschluss daran werden Aufbau und Ablauf der Testverläufe dargestellt. Abschließend erfolgt eine zusammenfassende Auswertung der verschiedenen Testläufe.

3.1 Datengrundlage und zentrale Erschließungsinstrumente in der sozialwissenschaftlichen Fachinformation

Die Datenbank SOLIS (Sozialwissenschaftliches Literaturinformationssystem) stellt die zentrale Literaturdatenbank der sozialwissenschaftlichen Fachinformation dar. Im Jahr 1980 eingerichtet, verzeichnet sie sozialwissenschaftliche Forschungsliteratur, die im deutschsprachigen Raum erschienen ist.[31] Aktuell umfasst SOLIS mehr als 400.000 Datensätze mit bibliographischen und inhaltlichen Angaben sowohl zu Monographien und Sammelwerken als auch zu Zeitschriftenaufsätzen, Sammelwerksbeiträgen und Grauer Literatur. Den größten Anteil des Bestandes liefert GESIS │Leibniz-Institut für Sozialwissenschaften selbst. Unter Verwendung der Titeldaten zu Monographien und Sammelwerken aus der Reihe A, herausgegeben von der Deutschen Nationalbibliothek, wird die relevante Forschungsliteratur bestellt und überwiegend von einem externen Dienstleistungsanbieter in Autopsie erschlossen. Zusätzlich werden knapp 300 Fachzeitschriften ausgewertet und auf Beitragsebene erschlossen. Daneben liefern Kooperationspartner einen Anteil des Bestandes. Hierzu zählen unter anderem das Institut für Arbeitsmarkt- und Berufsforschung (IAB) sowie das Wissenschaftszentrum Berlin für Sozialforschung (WZB).[32]

[31] Die ältesten Bestände reichen bis 1945 zurück. Sofern die Verlage im deutschsprachigen Raum angesiedelt sind, werden in SOLIS auch englischsprachige Titel verzeichnet. Generell liegen sämtliche Titeldaten sowohl auf Deutsch als auch auf Englisch vor.

[32] Zu den weiteren Datenlieferanten zählen unter anderem das Zentrum für Psychologische Information und Dokumentation (ZPID) sowie das Deutsche Institut für internationale päda-

Zentrale Instrumente der Inhaltserschließung

Die inhaltliche Auswertung der Literatur erfolgt unabhängig von der Dokumentart auf vier Ebenen. Dazu gehören das Kurzreferat, die Vergabe von sowohl inhaltlichen als auch methodischen Schlagwörtern bzw. Deskriptoren sowie die Klassifikation. Im Folgenden werden die einzelnen Regeln für die Literaturdokumentation in den Sozialwissenschaften vorgestellt.

Das *Kurzreferat* soll die zentralen Inhalte eines Dokuments knapp wiedergeben. Hierzu gehören in erster Linie sowohl theoretische und methodische als auch zentrale Aussagen zu den (empirischen) Befunden und Ergebnissen sowie die Schlussfolgerungen. Favorisiert wird somit das informative Referat, bei dem auch auf die im Dokument aufgeführten Resultate eingegangen wird. Daneben sollen „wichtige Nebenereignisse und Randaussagen" ebenso mit einfließen wie der behandelte Zeitraum und der untersuchte geographische Raum (vgl. GESIS 2005: 2). Je nach Komplexität und Vielschichtigkeit des Originaldokuments, so ist leicht nachvollziehbar, ist die Behandlung von Randthemen unter Umständen nur eingeschränkt möglich.

Das Kurzreferat ist als wesentliche Entscheidungsgrundlage für die Datenbanknutzer gedacht, ob das entsprechende Dokument – ohne dass es direkt vorliegt – für sie relevant ist, oder nicht. Dabei ist leicht nachvollziehbar, dass nicht bei jedem Dokument aus Gründen der Kürze sämtliche der einzelnen oben aufgeführten Bestandteile in das Referat aufgenommen werden können. Der angestrebte Umfang des Kurzreferats kollidiert somit unter Umständen mit den Erstellungsgrundsätzen. Unabdingbar ist jedoch die Regel, dass sämtliche der Schlüsselbegriffe eines Dokuments in das Kurzreferat einfließen sollen. Dieser Grundsatz dient dem Retrieval. Über die Freitextsuche können Dokumente über diese zentralen Begriffe gefunden werden.

Im Einzelnen gelten die Merkmale, wie sie in der DIN-Norm 1426 zur Erstellung von Inhaltsangaben in Information und Dokumentation genannt werden.

---

gogische Forschung (DIPF). Die Erschließung erfolgt unter Anwendung des Thesaurus sowie der Klassifikation Sozialwissenschaften. In Einzelfällen, wie etwa bei Datenlieferungen des ZPID werden hierfür Konvertierungslisten verwendet. Daneben bezieht GESIS │Leibniz-Institut für Sozialwissenschaften Rezensionsexemplare von der Zeitschrift für Politikwissenschaft sowie der Soziologischen Revue, die ebenfalls inhaltlich erschlossen und in die Datenbank aufgenommen werden.

Als Nutzergruppe wird dabei ein Fachpublikum angenommen. Zu den Merkmalen zählen sowohl Vollständigkeit, Genauigkeit und Objektivität als auch Kürze und Verständlichkeit. Die einzelnen Punkte stehen zum Teil in einem Spannungsverhältnis zueinander. Inwieweit die einzelnen Grundsätze tatsächlich umgesetzt werden, ist somit von der indexierenden Person abzuwägen. Zur Verständlichkeit gehört, dass vom Autor neu eingeführte Begriffe angeführt und erläutert werden sollen. Auch diese Begriffe sind somit über eine Freitextsuche auffindbar. Im Einzelnen orientieren sich Kurzreferate entlang einer bestimmten Gliederung. Wie streng diese eingehalten wird, ist erneut von der Dokumentengrundlage abhängig. Hierzu gehören die Darstellung von Thema, Vorgehensweise und Ergebnissen ebenso wie die Erwähnung von Schlussfolgerungen und Anwendungsbereichen (vgl. die ausführlicheren Hinweise in GESIS │Informationszentrum Sozialwissenschaften 2005).

Zulässig ist auch die Übernahme von Passagen aus dem Originaldokument, sogenannte Autorenreferate, die relevante Informationen zur Publikation enthalten. So eignen sich hierfür vornehmlich Textpassagen aus Zusammenfassungen und Schlussbetrachtungen sowie Abschnitte aus der Einleitung, die über die Zielsetzung und zentrale Hypothesen Auskunft geben.

Das zentrale Erschließungsinstrument bildet der Thesaurus Sozialwissenschaften. In ihm werden die Beziehungen zwischen den Begriffen und damit das begriffliche Umfeld der Deskriptoren, bestehend aus weiteren, engeren und verwandten Begriffen, abgebildet.[33] Verbunden mit der weit gefassten Definition des Gegenstandsbereichs der Sozialwissenschaften und der Möglichkeit zur Kompositabildung im Deutschen, weist er eine hohe Anzahl an Schlagwörtern auf.[34] So zählt der aktuelle Thesaurus circa 11.600 Einträge,

[33] Für eine Übersicht der Begriffsrelationen, die im Thesaurus Sozialwissenschaften enthalten sind, siehe Anhang Tab. 10.

[34] Die Sozialwissenschaften decken in Anlehnung an die UNESCO-Definition (1978) folgende Wissensgebiete und Anwendungsbereiche ab: Soziologie, Politikwissenschaft, Methoden der Sozialforschung, Sozialpsychologie, Bildungsforschung, Erziehungswissenschaft, Frauenforschung, Kommunikationswissenschaft, Ethnologie, Sozialgeschichte, Demographie, Sozialpolitik, Sozialwesen, Arbeitsmarkt- und Berufsforschung sowie sozialwissenschaftliche Aspekte der Psychologie, Wirtschaftswissenschaft, Umweltforschung, Rechtswissenschaft, Medizin und Gerontologie. Daneben wurden Bezeichnungen aus den Geistes- und Naturwissenschaften sowie Terme aufgenommen, die zum Gegenstandsbereich der Sozialwissenschaften zählen, auch wenn sie keine wissenschaftlichen Begriffe darstellen (vgl. IZ 2006).

davon rund 7.750 Deskriptoren und 3.850 Nicht-Deskriptoren (vgl. Informationszentrum Sozialwissenschaften (IZ) 2005). Bei Fragen zum fachlichen Scope und zum Spezialisierungsgrad der Schlagwörter bildet die Datenbank SOLIS einen entscheidenden Orientierungsrahmen. Um den Umfang des Vokabulars überschaubar zu halten und sprachliche Veränderungen leichter darzustellen, wurde konsequent das Prinzip der Postkoordination verfolgt: zusammengesetzte Begriffe werden (semantisch) zerlegt und durch die Aneinanderreihung von Einfachbezeichnungen repräsentiert.

Mit dem weit gefassten Gegenstandsbereich der Sozialwissenschaften ist eine weitere Schwierigkeit verbunden. So weisen manche Bezeichnungen je nach Fachteilgebiet bzw. -kontext unterschiedliche Bedeutungen auf. In diesen Fällen wird die Bedeutung der Bezeichnung entweder auf einen Kontext eingeschränkt, oder es wird auf andere eindeutige Deskriptoren verwiesen (vgl. IZ 2006).[35] Verweise auf unterschiedliche Deskriptoren finden sich in einzelnen Fällen auch je nach zeitlichem Kontext. Dies gilt etwa für geographische Begriffe. Dazu gehören Namen von Kontinenten und Teilkontinenten, Ländern und Regionen sowie für die Bundesrepublik Deutschland und Österreich von Bundesländern und für die Schweiz von Kantonen. So wird bei dem Deskriptor „Bundesrepublik Deutschland“ zusätzlich zwischen „Deutschem Reich“ und „Deutschland“ unterschieden. Eine weitere spezielle Regel im Umgang mit geographischen Begriffen bildet das geographische Upposting für außereuropäische Regionen.[36]

Die Klassifikation dient vorrangig der Zuordnung einer Publikation zu einem Wissenschaftsgebiet. Sie enthält 159 Klassifikationsnotationen aus etwa 20 Fachteilgebieten. Aufgeführt werden ausschließlich jene Gebiete, deren Methoden und Konzepte im Dokument angewendet werden. Als Richtwert gilt die Vergabe von bis zu drei Klassifikationsnotationen, inklusiver einer Hauptklassifikation. Daneben findet sich darin die Anweisung, sich im Zwei-

[35] Ein anschauliches Beispiel von Mehrdeutigkeit, die durch den Verweis auf alternative Deskriptoren gelöst wird, stellt die Bezeichnung Geschichte dar. Im Einzelnen gelten die Vorzugsbenennungen „historische Entwicklung“ im Sinne von Vergangenheit und Entwicklung, „Geschichtsunterricht“ sofern das Unterrichtsfach Geschichte gemeint ist, und „Geschichtswissenschaft“ für die wissenschaftliche Disziplin.

[36] Für eine Publikation zu Nepal werden dadurch automatisch etwa zusätzlich die Deskriptoren „Asien“ und „Südasien“ sowie „Entwicklungsland“ generiert.

felsfall für die Vergabe einer Klassifikationsnotation zu entscheiden (vgl. ebd.).

Als eine besondere Form der Erschließung gehört die abschließende Vergabe methodischer Deskriptoren. Damit wird beschrieben, wie der Autor das Thema behandelt. Im Anschluss an die Vergabe eines Methodendeskriptors zum Schwerpunkt der Arbeit können weitere Deskriptoren aufgeführt werden, bis das methodische Vorgehen, wie es im Dokument beschrieben wird, vollständig erfasst ist.

Titel	Migrationserfahrung - Fremdheit - Biografie : zum Umgang mit polarisierten Welten in Ost-West-Europa - Migration experience - foreignness - biography : dealing with polarized worlds in Eastern and Western Europe
Person(en)	Autor: Breckner, Roswitha (Universität Wien, Fak. für Sozialwissenschaften, Institut für Soziologie, Rooseveltplatz 2, Wien, Österreich)
Jahr	2009
Quelle	Wiesbaden: VS Verl. für Sozialwiss. 2009, 449 S. (Abb.) Serientitel: Forschung Gesellschaft ISBN: 978-3-531-16851-7
Inhalt	Die Autorin untersucht die biografische Bedeutung von Migrations- und Fremdheitserfahrungen. Diese werden anhand der Lebensgeschichten von Migrantinnen aus Rumänien, Ungarn, Polen und Russland, die vor 1989 in den Westen kamen, rekonstruiert. Folgende Fragestellungen stehen im Vordergrund: Gestalten sich Migrationserfahrungen als eigenständiger Erlebnis- und Erfahrungszusammenhang? Wie wird auf Migrationserfahrungen in der biographischen Konstruktion Bezug genommen? Wie verändert sich die Bezugnahme im Laufe der Lebensgeschichte bzw. in verschiedenen Kontexten? Welche biographischen Organisationsprinzipien von Erfahrungen, einschließlich ihrer Bezugschemata, werden durch migrationsspezifische Erfahrungen berührt? Welche Rolle spielen Fremdheitspositionen und -erfahrungen im Migrations- wie im biographischen Zusammenhang? Im empirischen Teil erfolgen hermeneutische Fallanalysen. Eine Besonderheit der Migrationsanalysen ist der Fokus auf die Ost-West-Problematik. (FR2)
Schlagwörter	Migration, postsozialistisches Land, Erfahrung, Biographie, Fremdheit, Bundesrepublik Deutschland, UdSSR-Nachfolgestaat, Lebenslauf, Europa, Osteuropa, Kalter Krieg, Polen, Rumänien, Russland, Ungarn, 20. Jahrhundert
Klassifikation	* 10304 - Migration
Methode	Befragung; biographische Methode; empirisch; empirisch-qualitativ; Theorieanwendung
Dokumenttyp	Monographie (gedruckt)
Sprache	deutsch
Produzent	GESIS - Leibniz-Institut für Sozialwissenschaften
Erfassungsnummer	20060112399
Copyright	GESIS - Datenbank SOLIS

Abbildung 2: Exemplarischer Datensatz aus der Datenbank SOLIS

3.2 Die Indexierungssoftware: Der MindServer

Die Testläufe für eine semiautomatische Erschließung der Datenbank SOLIS wurden mit der Indexierungssoftware MindServer der Firma Recommind durchgeführt. Diese Software wurde für die automatische Verschlagwortung und Klassifizierung nicht-erschlossener Dokumenteinheiten entwickelt. Dabei stellt die Software ein lernendes Verfahren dar. Anhand eines Trainingskorpus lernt das System, in welcher Form die Dokumente zu ordnen sind. Aus der Verarbeitung der Trainingsdokumente errechnet die Software Wahrscheinlichkeiten, nach denen einem neu zu erschließenden Dokument Kategorien, etwa in Form von Schlagwörtern und Klassifikationsnotationen, zugeordnet werden.[37] Die Software folgt somit in erster Linie einem statistischen Verfahrensansatz. Lediglich bestimmte Grundkomponenten, wie etwa Grundformreduktion und Derivation basieren auch auf computerlinguistischen Analysen. Dadurch, dass der Software zur Repräsentation des Dokumentinhalts ausschließlich Deskriptoren aus dem Thesaurus sowie die Klassifikation Sozialwissenschaften zur Verfügung gestellt werden – es handelt sich somit um ein Additionsverfahren – , wird das statistische Vorgehen zusätzlich um einen begriffsorientierten Verfahrensansatz erweitert.

Die Grundlage dieses automatischen Indexierungsverfahrens stellen die probabilistische latente semantische Indexierung sowie das Verfahren der Support Vector Machine dar. Beide Vorgehensweisen werden im Folgenden in ihren Grundzügen dargestellt.[38]

Die Funktionsweise der latenten semantischen Analyse verdeutlicht das Vektorraummodell von Gerald Salton (1987). Alle dokumentarischen Bezugseinheiten einer Dokumentensammlung, nach denen gesucht werden kann, werden darin als n-dimensionale Vektoren in einem n-dimensionalen Raum abgebildet. Die Gesamtzahl der Dimensionen ergibt sich dabei aus der Menge

[37] Dieser Lernvorgang erfolgt über den sogenannten Taxonomie-Browser. Auf dieser Basis können dem System im Rahmen des alltäglichen Geschäftsgangs neue Dokumente zur Indexierung zugeführt werden. Erscheint die Deskriptorenvergabe bei den neu zugeführten Dokumenten fehlerhaft, können diese im sogenannten Annotationstool korrigiert werden. Hierbei ist auch die Definition fester Indexierungsregeln möglich, die bei der weiteren Indexierung berücksichtigt werden.

[38] Die nachfolgenden Ausführungen sind sowohl in Bezug auf Aufbau und Formulierungen als auch Auswahl der Abbildungen sehr eng an Keil/Tiesler et al. 2010 angelehnt.

der zugelassenen Terme. Die einzelnen Werte der Vektoreneinträge werden neben weiteren Faktoren in erster Linie aus den Häufigkeiten der Terme pro Dokument berechnet. Sie bilden die sogenannten Termgewichte. Suchanfragen werden in dem Modell ebenso als Vektoren dargestellt. Auf diese Weise können Ähnlichkeiten zwischen den Dokumenten und den Anfragen ermittelt werden. Dieser Dialog wird in der nachfolgenden Abbildung schematisch dargestellt.

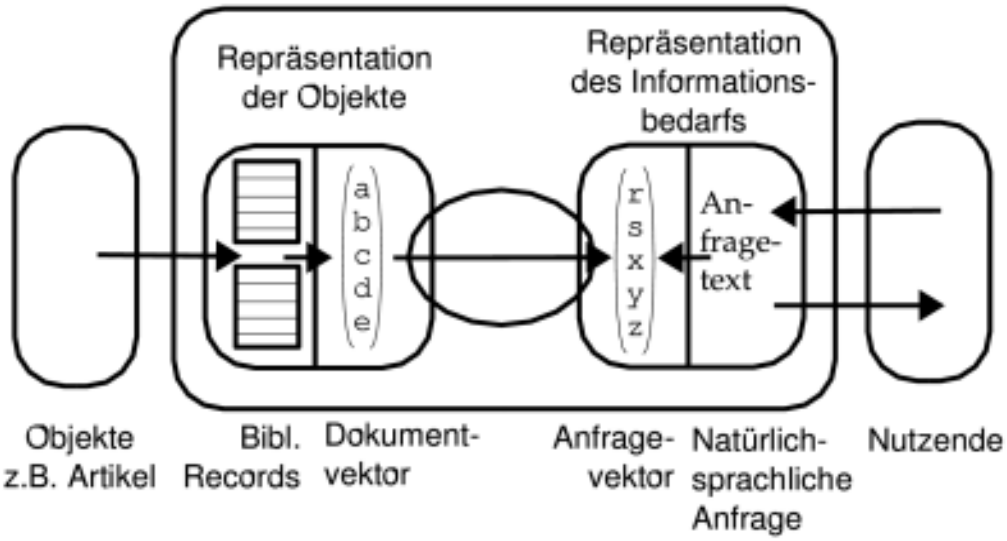


Abbildung 3: Schematische Darstellung eines Vektorraum-Text-Retrieval-Systems (Ferber 2003)

Der Objektbestand wird innerhalb des Systems sowohl durch die Metadaten, wie Deskriptoren und Klassifikation, als auch durch die einzelnen Wörter, die dem System etwa aus dem Abstract zu dem Dokument zur Verfügung stehen, abgebildet. Aus den verschiedenen Termgewichten derjenigen Terme, die zum Retrieval zugelassen sind, wird ein Dokumentvektor erstellt. In gleicher Weise wird mit den Termen verfahren, die jeweils in eine Suchanfrage eingehen, wodurch die Ähnlichkeit der Vektoren bestimmt werden kann. Hierfür wird häufig das Skalarprodukt der Vektoren oder das Cosinus-Maß verwendet. Durch die Ermittlung von Vektoren für jedes Dokument lässt sich schließlich die Ähnlichkeit zwischen den einzelnen Dokumenten berechnen (vgl. Keil/Tiesler et al. 2010).

Deerwester et al. (1990) unterscheiden in diesem Zusammenhang zwischen der Ein-Faktoren- und der Zwei-Faktoren-Analyse. Erstere basiert auf den Beziehungen zwischen den Dokumenten. Zweite schließt zusätzlich Bezie-

hungen zwischen Termen und Dokumenten mit ein. Diese Zwei-Faktoren-Analyse lässt sich an dem Beispiel einer Term-Dokument-Matrix (siehe Abb. 4) verdeutlichen. Dabei können Ähnlichkeiten zwischen Dokumenten hergestellt werden, obgleich im Einzelnen in ihnen nicht dieselben Begriffe aufgeführt sind.

Titles

c1	*Human* machine *interface* for Lab ABC *computer* applications
c2	A *survey* of *user* opinion of *computer system response time*
c3	The EPS *user interface* mangement *system*
c4	System and *human system* engineering testing of EPS
c5	Relation of *user*-perceived *response time* or error measurement
m1	The generation of random, binary, unorderd *trees*
m2	The intersection *graph* of paths and *trees*
m3	*Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4	*Graph minors*: A *survey*

Terms	**Documents**								
	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

Abbildung 4: Beispiel einer Term-Dokument-Matrix (Deerwester et al. 1990)

Abbildung 4 zeigt eine solche Term-Dokument Matrix. In den Zeilen sind die absoluten Häufigkeiten der Terme aufgeführt, die einzelnen Dokumente hingegen werden in den Spalten wiedergegeben. Entlang der Spalten lassen sich derart die Dokumentvektoren für jedes Dokument entnehmen. Bei der Zwei-Faktoren-Analyse werden nun zusätzlich die Beziehungen zwischen den Termen und den Dokumenten berücksichtigt und in die nachfolgenden Be-

rechnungen integriert.[39] Es zeigt sich, dass bestimmte Terme, die gar nicht in einem Dokument aufgeführt sind, für dieses gleichwohl einen hohen Wert zugesprochen bekommen können, da das Dokument anderen Dokumenten, die diesen Term sehr wohl enthalten, ähnlich ist. In Abbildung 5 zeigt sich dies exemplarisch für den Term „trees“ unter Dokument m4. Durch die Ähnlichkeit des Titels zum Dokumenttitel von m3 erhält dieser Term einen hohen Wert, obgleich er nicht im Titel von m4 aufgeführt ist.

Terms	**Documents**								
	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40.	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	**0.66**
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

Abbildung 5: Rekonstruierte Term-Dokument-Matrix (Landauer et al. 1998)

In der MindServer Software der Firma Recommind wird durch das Verfahren der sogenannten prohabilistischen latenten semantischen Analyse, einem patentierten Algorithmus, eine derartige statistische Analyse von gemeinsamen Wortvorkommen in Dokumenten vorgenommen. Sich wiederholende Kontexte, Themen oder Konzepte, in denen jeweils eine bestimmte Gruppe von Wörtern vorkommen, werden so identifiziert (vgl. Puzicha 2009, zitiert nach Keil/Tiesler et al. 2010: 8). Die Information über das Auftreten eines Terms in einem bestimmten Dokument wird gespeichert und für die weitere Indexierung genutzt. Die Erweiterung der latenten semantischen Analyse besteht in der Einführung einer Variablen, die die Wahrscheinlichkeit angibt, mit der ein Term oder ein Dokument zu einer bestimmten Klasse gehört, dabei werden auch bestimmte häufig vorkommende Wortmuster extrahiert (vgl. Holl 2009). Diese Klasse entspricht einem bestimmten Kontext bzw. Thema oder

[39] Einen zentralen Bereich dieser Berechnungen bildet die Singulärwertzerlegung (Singular Value Decomposition). Aus der Ursprungsmatrix werden einzelne Teilmatrizen gebildet. Sehr geringe Werte werden dabei ausgeschlossen. Nach weiteren Operationen werden diese Teilmatrizen erneut wieder zusammengeführt.

Konzept. Auf der Grundlage der Trainingsmenge leitet die Software Wahrscheinlichkeiten ab, mit denen ein Wort in einem Dokument zu einer bestimmten Klasse gehört. Dieser Lernschritt wird für die Anordnung der Terme und Dokuments im Vektorraum genutzt.

Schließlich basiert die MindServer Software auf einer sogenannten Support Vector Machine (SVM). Diese bildet einen Klassifikator, der dazu eingesetzt wird, den vieldimensionalen Vektorraum in Form von linearen Ebenen zu trennen, um dadurch Klassenzugehörigkeiten zu bestimmen. Lassen sich derartige lineare Ebenen zunächst nicht ziehen, erweitert die SVM den Vektorraum um die erforderliche Anzahl zusätzlicher Dimensionen. Die neu berechnete lineare Ebene wird dann „in den ursprünglichen Vektorraum rücktransformiert“ (Keil/Tiesler et al. 2010).

Gleichwohl handelt es sich bei der semantischen Analyse „lediglich“ um ein statistisches Verfahren. Die semantischen Strukturen innerhalb der Dokumente werden nicht erkannt. Es lässt sich somit nicht von einem Verstehen der Dokumente durch das System sprechen.[40] Allerdings geht die Analyse über die Verarbeitung der einzelnen Wörter als Zeichenkette hinaus. Durch den Abgleich der einzelnen Dokumentdaten mit dem Gesamtbestand der Dokumentenkollektion innerhalb des gesamten Vektorraums wird versucht, auch diejenigen Begriffe bzw. Begriffsinhalte zu erfassen, die lediglich latent in der Dokumentgrundlage enthalten sind. Diese Berücksichtigung des (weitergefassten) Kontextes, in dem ein Begriff vorkommt, ist vor allem für implizite Inhalte, die nicht lexikalisch ausgedrückt werden, von großem Vorteil für das Retrieval (vgl. Betram 2005).

3.3 Aufbau, Ablauf und Auswertung der Testläufe

Im Rahmen der vorliegenden Arbeit wurde die Indexierungssoftware MindServer der Firma Recommind entlang von vier Testläufen getestet. Im Folgenden wird zunächst auf die Auswahl der Stichprobe sowie den Aufbau der

[40] Der Softwareanbieter Recommind wirbt gleichwohl damit, dass diese Software auf dem gleichen Niveau erschließe wie intellektuelle Indexierer. Dabei beruft sich das Unternehmen auf unabhängige Tests des Fraunhofer-Instituts (vgl. Recommind 2011).

Testläufe eingegangen, bevor im Anschluss die wesentlichen Auswertungskriterien vorgestellt werden.

Auswahl und Beschreibung der Stichprobe

Zur Evaluation der Indexierungssoftware wurde zunächst eine Stichprobe von 280 Dokumenten aufgebaut. Hierbei handelte es sich um Dokumente, die zwischen August 2009 und August 2010 in die Datenbank SOLIS aufgenommen wurden. Die Grundlage der automatischen Indexierung bildeten sowohl der Dokumenttitel als auch der Abstract-Text. Bei der Auswahl der Dokumente wurde darauf geachtet, dass die Anteile der Dokumentarten dem Zuwachs der Datenbank ab dem Erscheinungsjahr 2005 entsprachen. Verbunden mit Empfehlungen, wie sie aus der vorletzten Evaluation des Instituts im Jahr 2004 hervorgingen (vgl. Leibniz Gemeinschaft – Der Senat, 2005), wurden ab diesem Zeitpunkt vermehrt Sammelwerksbeiträge in die Datenbank SOLIS aufgenommen.

Im Einzelnen handelt es sich bei 34 Prozent um Zeitschriftenaufsätze, bei 36 Prozent um Sammelwerksbeiträge, bei 22 Prozent um Monographien sowie bei acht Prozent um Gesamtaufnahmen von Sammelwerken. Die Indexate der Gesamtaufnahmen von Sammelwerken weisen dabei zwei Spezifika auf. Zum einen handelt es sich bei den Inhaltstexten zumeist um Autorenreferate. Aktuell werden für die Gesamtaufnahmen der Sammelwerke keine Abstracts mehr erstellt. Zum anderen wird bei den Gesamtaufnahmen von Sammelwerken im Anschluss an das Autorenreferat das Inhaltsverzeichnis mit aufgenommen.

Diese unterschiedlichen Abstract-Arten wurden bei der Zusammensetzung der Stichprobe ebenfalls berücksichtigt. So befinden sich zum einen in der Datenbank SOLIS Autorenreferate sowie Textteile, die bei der intellektuellen Erschließung aus der Dokumentvorlage übernommen wurden. Der Anteil an der Gesamtmenge beträgt etwa 40 Prozent. Den restlichen Teil bilden Abstracts, die entsprechend dem Regelwerk für die Literaturdokumentation Sozialwissenschaften, wie sie in starker Anlehnung an DIN 1426 (Inhaltsangaben in Information und Dokumentation) formuliert wurden, neu erstellt wurden. Den Hintergrund bildeten Überlegungen, aus Kostengründen den Anteil an Dokumenten, zu denen ein Abstract geschrieben wurde, mittel- und

langfristig zurückzufahren. So sollen perspektivisch vermehrt Textpassagen, etwa aus Einleitung und Klappentext, übernommen werden. Die Evaluation des MindServer geht somit auch der Frage nach, inwiefern sich diese Textgrundlagen für eine computergestützte Indexierung eignen. In die Stichprobe sind 164 Dokumente eingegangen, deren Abstract nach dem Regelwerk erstellt wurden, und 116 Dokumente mit Autorenreferat.[41]

Beschreibung und Auswertung der Testläufe

Für die vorliegende Evaluationsstudie wurden vier Testläufe durchgeführt. Diese bauten aufeinander auf und variierten sowohl hinsichtlich der vorgenommenen Systemeinstellungen als auch in Bezug auf die zu Grunde gelegten Trainingsmengen.[42] Verbunden mit den jeweils gewählten Testeinstellungen werden zunächst die Testläufe I und II und im Anschluss die Testläufe III und IV hinsichtlich der erzielten Indexierungsergebnisse miteinander verglichen. Im Anschluss werden schließlich beide Testblöcke (I und II sowie III und IV) miteinander verglichen.

Die Einstellungsveränderungen betrafen bei jedem der Testläufe sowohl die Vergabe von Deskriptoren als auch die Zuordnung der Klassifikation. Verändert wurden jeweils Minimum und Maximum der Trainingsdokumente pro Kategorie sowie Minimum und Maximum der generierten Kategorien pro Dokument, das Verhältnis zwischen *Recall* und *Precision* und die Textlänge, die erforderlich war, um überhaupt ein Indexierungsergebnis zu erzielen.[43]

Die zentralen Größen bei der *Auswertung* der Testläufe bilden die Standardmaße zur Evaluation des Information-Retrieval: *Recall* und *Precision* (siehe

[41] Im Verlauf der Tests variierte die Gesamtzahl der Dokumente in der Stichprobe mehrfach. Dies hatte verschiedene Gründe. Zum einen stellte sich bei der Durchsicht der Abstract-Arten heraus, dass sechs Dokumente lediglich ein englischsprachiges Autorenreferat und ein Dokument ausschließlich ein französischsprachiges Inhaltsverzeichnis aufwiesen. Zum anderen kam es vor, dass die Indexierungssoftware keinerlei Deskriptoren oder Klassifikationen generierte. Dies hing teilweise mit den vorgenommenen Systemeinstellungen für die verschiedenen Testläufe zusammen. Im Einzelnen wird die genaue Gesamtzahl an Dokumenten der Stichprobe jeweils angegeben.

[42] Zur Gesamtdarstellung der Testeinstellungen siehe im Anhang Tab.11.

[43] Die Bezeichnung Kategorie meint in diesem Zusammenhang sowohl Deskriptor (bezogen auf den Thesaurus Sozialwissenschaften) als auch Klassifikationsnotation (bezogen auf die Klassifikation Sozialwissenschaften).

Kap. 2.2).[44] Daneben werden die *Indexierungsbreite* der automatisch generierten Indexierungsergebnisse sowie die *Indexierungskonsistenz* ermittelt. Die Gesamtanzahl der automatisch vergebenen Deskriptoren wird dabei noch einmal differenziert betrachtet. So wird zusätzlich zur Menge der Deskriptoren, die ebenso intellektuell vergeben wurden, einerseits zwischen inhaltlich fehlerhaften – im Sinne von inhaltlich unzutreffenden bzw. irreführenden Indexierungstermen – sowie andererseits zusätzlich passenden Deskriptoren unterschieden.[45]

In ähnlicher Weise wird auch das Klassifizierungsergebnis differenziert betrachtet. Neben der Unterscheidung zwischen inhaltlich unzutreffenden sowie zusätzlich passenden Klassifikationen wird hierbei gesondert festgehalten, ob und auf welcher Position des Ranking nach dem Konfidenzwert die intellektuell vergebene Hauptklassifikation ermittelt wurde.

Im Einzelnen wurden zur Beurteilung der Indexierungsergebnisse durch den MindServer folgende *Variablen* erfasst:

- Abstract-Art (informatives Abstract oder Autorenreferat),
- Textlänge des Abstracts[46],
- Anzahl der vergebenen Deskriptoren und Klassifikationen von MindServer sowie der Dokumentart (Monographie, Zeitschriftenartikel, Sammelwerksbeitrag oder Gesamtaufnahme eines Sammelwerks),
- intellektuellen Erschließung und Ermittlung der Überschneidungsmenge,

[44] Im Rahmen der vorliegenden Arbeit wird somit vornehmlich eine Teilkomponente des Informationssystems, nämlich das Erschließungsverhalten automatischer Indexierungsverfahren, betrachtet. Der Blick ist dabei im Wesentlichen auf die Dokumentebene gerichtet. Der Evaluierungsansatz besteht in erster Linie darin, die Indexierungsergebnisse automatischer Sacherschließungsansätze denen der intellektuellen Erschließung gegenüberzustellen. Die manuellen Indexierungen wurden als Vergleichsdaten herangezogen. Untersucht wird, inwiefern sich intellektuelle und automatische Sacherschließung einander annähern. Relevant erscheint ein Indexierungsterm folglich in erster Linie dann, wenn er im intellektuell generierten Indexat aufgeführt wird.

[45] Auf weitere Kategorien des Kriteriensets nach Stock/Stock (2008) einzugehen, wie Indexierungstiefe (bestehend aus Indexierungsbreite und -spezifität) sowie Indexierungseffektivität und -konsistenz, war im Rahmen dieser Arbeit aus Zeitgründen nicht möglich.

[46] Der Einfluss der Textlänge des Abstracts auf das Indexierungsergebnis wurde ausschließlich während des ersten Testlaufs evaluiert. Die Zeilenanzahl wurde entsprechend der Anzeige des Abstracts bzw. Autorenreferates (z.T. inklusive Inhaltsverzeichnis) in aDIS berechnet.

- Ermittlung von Deskriptoren und Klassifikationen, die ausschließlich von MindServer vergeben wurden und vom Bearbeiter der Untersuchung ebenfalls als relevant angesehen wurden,
- Erfassung fehlender Titelbegriffe,
- Erfassung „falscher“ Deskriptoren und Klassifikationen, die von MindServer vergeben wurden. Hierbei wurde unterschieden, ob es sich dabei um sachlich oder geographisch „falsche“ Deskriptoren handelte.[47]
- Erfassung, ob bei den Klassifikationen, die von MindServer vergeben wurden, die Hauptklassifikation gefunden wurde. Es wurde aufgeführt, ob MindServer die Hauptklassifikation gefunden hat und auf welchem Rang sie entsprechend dem Konfidenzwert platziert wurde,
- Ermittlung der Indexierungskonsistenz sowohl für die Vergabe der Deskriptoren als auch für die Zuordnung der Klassifikationsnotationen.[48]

Sämtliche dieser Variablen wurden für jedes Dokument in einer Tabelle aufgeführt.[49]

3.3.1 Die Testläufe I und II

Für die ersten beiden Testläufe wurde der MindServer mit circa 368.000 Dokumenten der Datenbank SOLIS trainiert. Dies entspricht dem Gesamtbestand der Datenbank bis Juli 2009. Alle bereits intellektuell vergebenen De-

[47] Diese Unterscheidung wurde über die gesamte Stichprobe hinweg ausschließlich bei der Auswertung des ersten Testlaufs vorgenommen. In einer Zwischenuntersuchung (vgl. Anhang) wurde diese Unterscheidung auch bei denjenigen Deskriptoren vorgenommen, die sowohl intellektuell als auch automatisch generiert wurden. Bei der Auswertung zeigte sich, dass die Überschneidungsmenge oftmals zahlreiche Geographika aufwies. Für spätere Testläufe wäre es sinnvoll, eine Unterscheidung zwischen sachlichen und geographischen Deskriptoren bei der Überschneidungsmenge über die gesamte Stichprobe hinweg vorzunehmen. Für die weiteren Testläufe (II bis IV) bedeutete eine derartige Auswertung einen zu hohen Zeitaufwand.

[48] Die Ergebnisse zur Indexierungskonsistenz werden erst in den Ergebnisteilen (Zwischenfazit sowie Kap. 3.4) in die Darstellung mit aufgenommen. Dieses Vorgehen wurde gewählt, um die Diagramme übersichtlich zu halten.

[49] Nach Auswertung des ersten Testlaufs wurden allerdings weder die Dokument-Datensätze aus aDIS noch die Ergebnisdarstellung des MindServers für jedes Dokument festgehalten. Dies bedeutete, dass bei allen bisherigen Analyseschritten (siehe weiter unten) zunächst für jedes der betrachteten Dokumente eine Abfrage in aDIS sowie dem MindServer durchgeführt und die bis dato von den Bearbeitern vorgenommenen Analysen zunächst rekonstruiert werden mussten. Dadurch erwies sich die Einarbeitung in die vorgenommene Bewertung der MindServer-Ergebnisse als sehr mühsam und zeitaufwändig.

skriptoren und Klassifikationen der trainierten Dokumente wurden von der Indexierungssoftware als eindeutig zutreffend (Konfidenzwert = 100 Prozent) eingestuft.

Für die anschließende inhaltliche Erschließung der Testdokumente wurden alle 7.750 Deskriptoren des Thesaurus Sozialwissenschaften sowie alle 167 Möglichkeiten der Klassifikation zugelassen.[50]

Testlauf I

Beim ersten Testlauf wurden die Standardeinstellungen des MindServer nicht verändert. Im Einzelnen bedeutete dies, dass zum einen keinerlei Einschränkung bzw. Festlegung bei der Anzahl der von MindServer vergebenen Deskriptoren und Klassifikationen vorgenommen wurde. Die Anzahl der generierten Deskriptoren und Klassifikationen konnte zwischen den Trainingsdokumenten sehr stark variieren. Zum anderen wurde keinerlei Einschränkung bei der Textlänge vorgenommen, die aus Titel und Abstract bzw. Autorenreferat zusammengenommen mindestens vorliegen musste, um einen Deskriptor bzw. eine Klassifikation zu generieren. Des Weiteren sahen die Standardeinstellungen vor, dass bereits ein Dokument mit der entsprechenden Kategorie in der Trainingsmenge ausreichte, um den entsprechenden Deskriptor bzw. die entsprechende Klassifikation automatisch zu vergeben. Die maximale Anzahl an Dokumenten, die pro Kategorie in der Trainingsmenge vorliegen durfte, betrug 25.000 Dokumente. Das Verhältnis zwischen *Recall* und *Precision* betrug +20.0. Damit wurde eine Indexierung zugunsten des *Recall* vorgenommen.[51]

Ergebnisse des ersten Testlaufs

Beim ersten Testlauf unterscheiden sich manuelle und automatische Erschließung deutlich in Bezug auf die *Indexierungsbreite*. 14 intellektuell vergebene Deskriptoren stehen durchschnittlich 19 maschinell generierten Deskriptoren

[50] Verbunden mit den Systemeinstellungen des zweiten Testlaufs reduzierten sich sowohl die Gesamtzahl der berücksichtigten Deskriptoren als auch die Anzahl der Klassifikationen. Aus systemtechnischen Gründen ließ sich die genaue Anzahl nicht ermitteln.

[51] Es wird davon ausgegangen, dass ein Verhältnis zwischen *Recall* und *Precision* zugunsten des *Recall* nicht grundsätzlich die Voreinstellungen des MindServer darstellen. Doch bildete dieses Verhältnis die Standardeinstellung, mit der die Software bei GESIS │Leibniz-Institut für Sozialwissenschaften betrieben wird.

gegenüber. Dies wird darauf zurückgeführt, dass kein *cut-off* Level festgelegt und der *Recall* um 20 Prozent höher eingestellt wurde als die *Precision*. Die Erkennung von Ähnlichkeiten zwischen dem zu indexierenden Dokument und den Dokumenten in der Trainingsmenge vollzieht sich gleichsam in einem weiter gefassten Kontextraum. Auch schwächer ausgeprägte Ähnlichkeiten zu Dokumenten werden registriert, was die durchschnittliche Anzahl maschinell generierter Deskriptoren erhöht.[52]

Im Durchschnitt sind sechs dieser 19 Deskriptoren identisch. Knapp fünf Deskriptoren sind im Durchschnitt inhaltlich fehlerhaft, d.h. sie liegen außerhalb des Themen- und Begriffsumfeldes der zugehörigen Publikation.[53] Hieraus ergeben sich ein *Precision*-Wert von 32,4 Prozent und ein *Recall*-Wert von 44,0 Prozent.

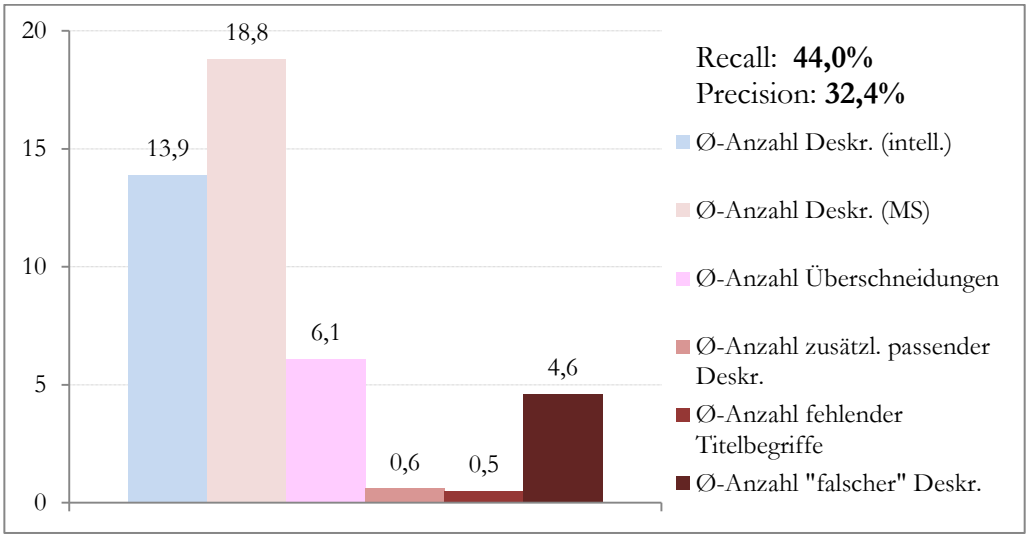


Abbildung 6: Vergabe der Deskriptoren Testlauf I (n=271)[54]

Aus der Verteilung der *Recall*- und *Precision*-Werte auf die Gesamtstichprobe geht hervor, dass in 46 Prozent der Fälle (bei 124 der 271 Dokumente) der

[52] Diese Erklärung wird auch durch die nachfolgenden Tests bestätigt. So werden etwa bei einem ausgeglichenen Verhältnis zwischen *Recall* und *Precision* im vierten Testlauf lediglich knapp 14 Deskriptoren vergeben. (Siehe im Einzelnen weiter unten).

[53] Bei viereinhalb dieser Deskriptoren handelt es sich um sachliche Deskriptoren. Bei durchschnittlich einem halben Deskriptor wurde eine fehlerhafte geographische Einordnung vorgenommen. Bei der Auswertung der nachfolgenden Testläufe wurde diese Differenzierung zwischen fehlerhaft vergebenen Sachschlagwörtern und Geographika nicht mehr vorgenommen.

[54] Die neue Gesamtzahl der Stichprobe ergab sich dadurch, dass bei sechs Dokumenten lediglich englischsprachige Abstracts, bei einem Dokument ein französischsprachiges Inhaltsverzeichnis vorlag und bei zwei Dokumenten die Indexierungssoftware keinerlei Deskriptoren vorschlug.

Recall-Wert bei 50 oder mehr Prozent liegt (siehe Abb. 2). Bei über einem Viertel der Dokumente (26 Prozent bzw. 72 der 271 Dokumente) beträgt der *Recall*-Wert sogar 60 oder mehr Prozent.

Ein anderes Bild ergibt sich bei der Verteilung der *Precision*-Werte. Hier liegen hingegen nur 20 Prozent (54 Dokumente) bei einem Wert von 50 und mehr Prozent. 80 Prozent der Dokumente liegen unterhalb der 50-Prozent-Marke. Für über die Hälfte der Dokumente (143 Dokumente) liegt der *Precision*-Wert zwischen 20 und 39 Prozent.

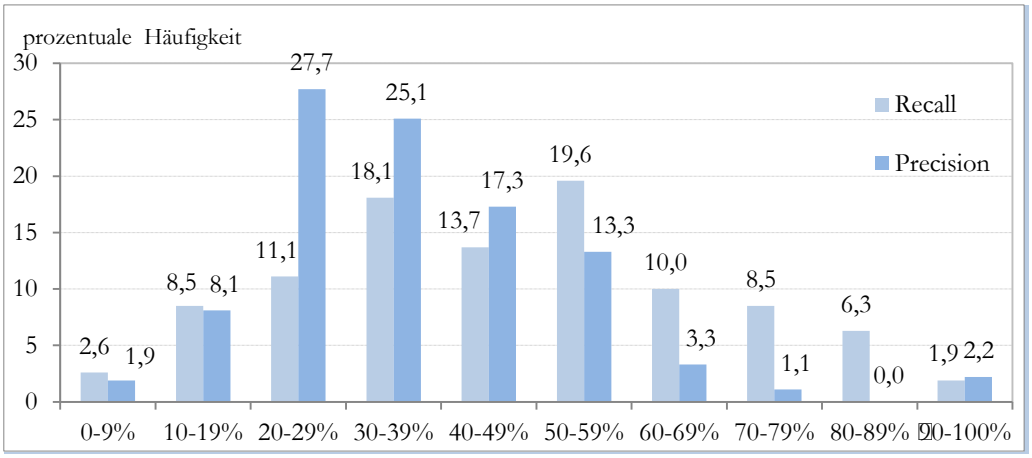


Abbildung 7: Verteilung der *Precision*- und *Recall*-Werte auf die Gesamtstichprobe Testlauf I (n=271)[55]

Insgesamt werden in 82 Prozent der Fälle (bei 222 der 271 Dokumente) inhaltlich fehlerhafte Deskriptoren vergeben. In knapp 39 Prozent der Fälle (bei 105 der 271 Dokumente) werden zusätzlich passende Deskriptoren, die intellektuell nicht ausgewählt worden waren, vergeben. Zählt man diese Deskriptoren noch zur Menge der relevanten Deskriptoren hinzu, so steigen *Recall*- und *Precision*-Wert auf 48,6 bzw. 36,4 Prozentpunkte.

Die eingangs erwähnte Indexierungsbreite lässt sich in Bezug auf die *Dokumentart* differenzieren (siehe Abb.8). Bei den Gesamtaufnahmen von Sammelwerken liegt die Anzahl der Deskriptoren mit durchschnittlich 14 vergebenen Deskriptoren deutlich unter dem Durchschnitt der Gesamtmenge mit

[55] Abbildung 7 liest sich beispielsweise folgendermaßen: Bei 27,7 Prozent der Dokumente aus der Gesamtstichprobe stimmen die von MindServer vergebenen Deskriptoren im Durchschnitt zu 20 bis 29 Prozent mit den intellektuell vergebenen Deskriptoren überein. Bei 18,1 Prozent der Dokumente aus der Gesamtstichprobe machte der Anteil der (gemeinsam) sowohl von MindServer als auch intellektuell vergebenen Deskriptoren durchschnittlich 30 bis 39 Prozent der von MindServer insgesamt vergebenen Deskriptoren aus.

knapp 19 Deskriptoren. Dies wird auf die Vielzahl an Themen und Aspekten zurückgeführt, die in manchen Sammelwerken behandelt werden und die durch die Angabe des Inhaltsverzeichnisses im Datensatz unmittelbar in den Indexierungsvorgang mit einfließt. Die Themenvielfalt erschwert die Erkennung von Ähnlichkeiten zu anderen Dokumenten aus der Trainingsmenge und beeinträchtigt hierdurch die eindeutige Zuordnung von Deskriptoren.

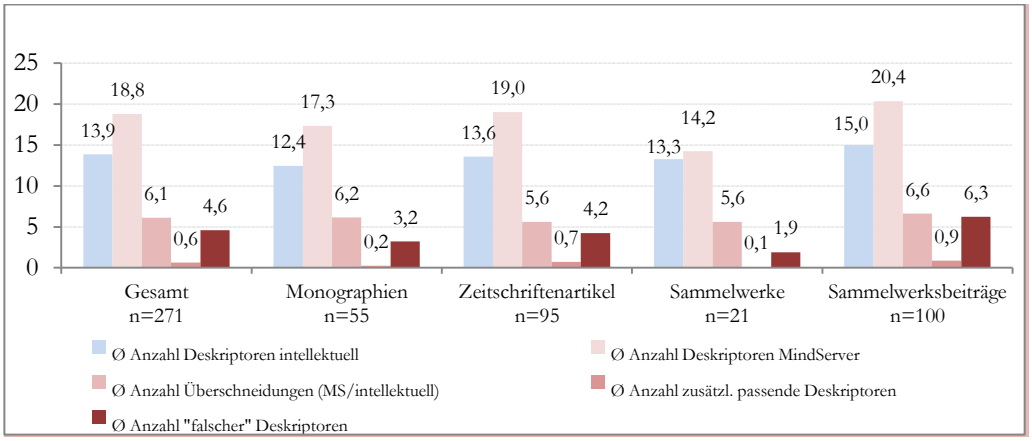


Abbildung 8: Durchschnittliche Anzahl vergebener Deskriptoren nach Dokumentart Testlauf I (n=271)

Auf der Beitragsebene fällt die Indexierungsbreite hingegen höher aus. Dies korreliert mit der intellektuell generierten Anzahl an Schlagwörtern für Sammelwerksbeiträge und Zeitschriftenartikel. Die deutlichere Eingrenzung eines Themas auf der Beitragsebene erleichtert es, Beziehungen zu Dokumenten aus der Trainingsmenge aufzubauen und daraus Deskriptorenvorschläge abzuleiten. Gleichwohl wird aus der relativ hohen Anzahl fehlerhafter Vorschläge sowohl bei Sammelwerksbeiträgen als auch bei Zeitschriftenartikeln die Schwierigkeit deutlich, die spezifische Eingrenzung eines Themas adäquat abzubilden.[56]

[56] Diese Erklärung wird durch die Ergebnisse der nachfolgenden Testläufe bestätigt. Der Aufbau fachteilgebietsspezifischer Trainingsmengen für die Testläufe III und IV führt dazu, dass sich sowohl für Sammelwerksbeiträge als auch für Zeitschriftenartikel die Anzahl der von der Indexierungssoftware im Durchschnitt vergebenen Deskriptoren um zwei Deskriptoren reduziert. Die Erschließung sehr spezifischer Themen scheint für die Indexierungssoftware auf der Grundlage fachteilgebietsspezifischer Trainingsmengen leichter möglich. Dies zeigt sich etwa in einem Vergleich der Testläufe I und III. Für beide Testläufe wurden die Standardeinstellungen des MindServer verwendet. Bei der maschinellen Erschließung von Sammelwerksbeiträgen für die Fachteilgebiete Soziologie, Politikwissenschaft und Geisteswissenschaften (n=43) werden im Durchschnitt 18,2 (statt 20,4) Deskriptoren generiert. Bei der Indexierung von Zeitschriftenartikeln (n=70) für die drei Fachteilgebiete werden durchschnittlich 17,3 (statt 19,0) Deskriptoren vergeben. Auch die fehlerhafte Vergabe von Deskriptoren lässt sich durch den Aufbau fachteilgebietsspezifischer Trainingsmengen deutlich reduzieren. Diese liegt beim dritten Testlauf für die Sammelwerksbeiträge bei 1,3 Deskripto-

Bei der *Klassifikation* erzielt die Indexierungssoftware ein besseres Ergebnis. Unter Verwendung der Standardeinstellungen werden von MindServer mit durchschnittlich vier Klassifikationen zwei mehr als bei der intellektuellen Erschließung vergeben (siehe Abb. 9). Die Überschneidung liegt bei durchschnittlich eineinhalb Klassifikationen. Bei rund 76 Prozent (bei 200 der 262 Dokumente) wird die Hauptklassifikation gefunden, allerdings liegt sie im Durchschnitt auf der zweiten Position des Rankings. Zusätzlich passende Klassifikationen werden bei etwa 18 Prozent (47 Dokumenten) vergeben, fehlerhafte Klassifikationseinordnungen bei knapp 63 Prozent (165 Dokumenten) vorgenommen.

Erneut unter Berücksichtigung der Dokumentart betrachtet, fällt auf, dass die Unterschiede bei der Klassifikation zwischen den Dokumentarten gering ausfallen. Ähnlich wie bei der Vergabe der Deskriptoren erzielen die Sammelwerksbeiträge die höchste Anzahl sowohl an Klassifikationen – dies entspricht auch der intellektuellen Vergabepraxis – als auch die meisten fehlerhaften Klassifikationszuordnungen. Dies wird erneut auf die hohe Spezifik der Dokumentinhalte zurückgeführt. Die Eingrenzung der Beiträge auf ein spezifisches Thema erleichtert einerseits die Ähnlichkeitserkennung zu anderen Dokumenten und damit die Klassifikationszuordnung. Andererseits steigt die Wahrscheinlichkeit einer fehlerhaften Einordnung, was in ähnlicher Weise auch für Zeitschriftenartikel gilt. Daneben zeigt eine stichprobenartige sehr genaue Durchsicht dieser Dokumentart, dass häufig eine größere Anzahl von Klassifikationsvorschlägen aus einer Klassifikationsgruppe kommt. Auch dies deutet darauf hin, dass es der Indexierungssoftware bei Verwendung der Standardeinstellungen Probleme bereitet, die spezifische Perspektive auf ein Themengebiet abzubilden. Die wenigsten fehlerhaften Zuordnungen nimmt die Indexierungssoftware bei Monographien und Gesamtaufnahmen von Sammelwerken vor. Gleichwohl erscheinen die Unterschiede insgesamt gering und die Anzahl dieser Dokumentarten an der Gesamtstichprobe um eini-

---

ren. Für die Zeitschriftenaufsätze beträgt sie 0,9 Deskriptoren. Gleichwohl soll auf die eingeschränkte Aussagekraft dieser Ergebnisse aufgrund der geringen Dokumentenanzahl hingewiesen werden. Sie zeigt sich auch darin, dass sich etwa bei der Eingrenzung auf die drei Fachteilgebiete bei den Zeitschriftenartikeln auch die durchschnittliche Anzahl der intellektuell vergebenen Deskriptoren um zwei Schlagwörter (von 13,6) auf 15,5 erhöhte.

ges niedriger, weswegen in diesem Zusammenhang keine weiteren Aussagen vorgenommen werden.[57]

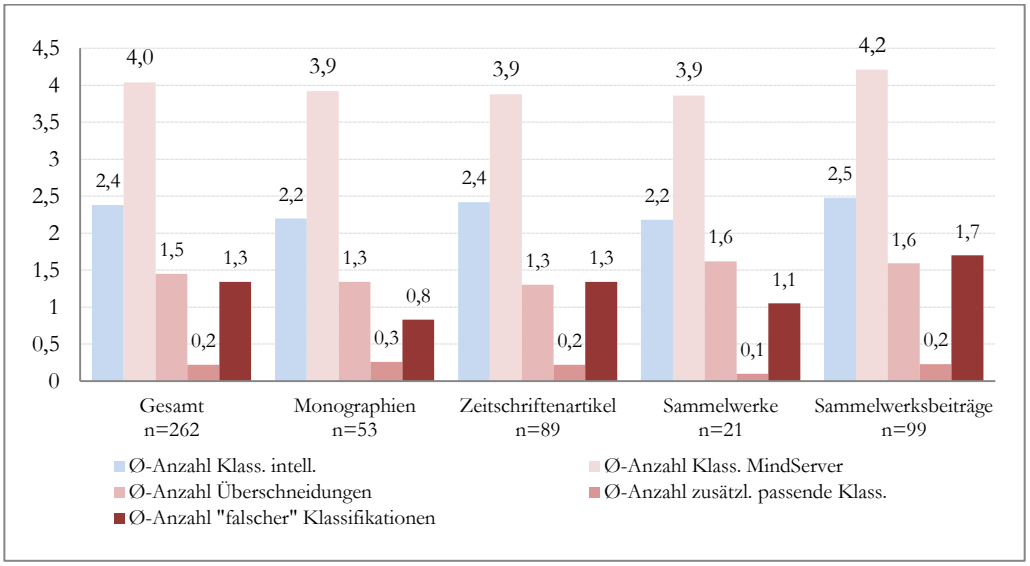


Abbildung 9: Durchschnittliche Anzahl vergebener Klassifikationen nach Dokumentart Testlauf I (n=262)[58]

Mit Blick auf die *Textlänge* zeigt sich, dass die Indexierungssoftware mit zunehmender Textlänge bessere Ergebnisse hervorbringt. Der Anteil der Überschneidungsmenge an der Gesamtzahl der maschinell generierten Deskriptoren nimmt konstant zu.[59] Der *Precision*-Wert steigt somit mit zunehmender Textlänge, während die Anzahl inhaltlich falsch vergebener Deskriptoren kontinuierlich sinkt.

	Anteil der Überschneidungsmenge an der Ergebnismenge MindServer (*Precision*-Wert)	**Ø-Anzahl "falscher" MindServer Deskriptoren**	**Anteil der Dokumente mit "falschen" Deskriptoren an der Gesamtmenge**

[57] Ein Grund hierfür könnte darin liegen, dass für beide Dokumentarten der Anteil von Autorenreferaten sehr hoch ist. Bei Durchsicht der Dokumentarten nach Materialart zeigte sich, dass 16 der 53 Monographien mit einem Abstract, hingegen 37 Dokumente mit einem Autorenreferat erschlossen worden waren. Gesamtaufnahmen von Sammelwerken werden aktuell sogar ausschließlich mit einem Autorenreferat versehen. In der Stichprobe wiesen 13 der 22 Gesamtaufnahmen ein Autorenreferat auf. Im weiteren Verlauf der Auswertung weisen die automatisch generierten Klassifikationsvorschläge bei Autorenreferaten minimal mehr Überschneidungen mit den intellektuell vorgenommenen Zuordnungen auf als bei Abstracts.

[58] Bei lediglich 262 der 271 Dokumente wurden von der Indexierungssoftware Klassifikationen vergeben, dabei stellten Sammelwerksbeiträge den größten Teil der nicht klassifizierten Dokumente. Dies wird auf den hohen Spezialisierungsgrad dieser Publikationen zurückgeführt, die vor allem aus den Randbereichen der Datenbank kamen. So wurden intellektuell bei diesen Dokumenten häufiger etwa die Klassifikationen „Philosophie, Ethik, Theologie" oder „Verwaltungswissenschaft" vergeben.

[59] Mit „Überschneidungsmenge" ist im Folgenden stets die Anzahl an Deskriptoren gemeint, die sowohl intellektuell als auch maschinell generiert wird.

Gesamt n=271	32%	4,8	82,5%
Textlänge 1 (0-10 Zeilen) n=51	26,9%	7,4	92,5%
Textlänge 2 (10-20 Zeilen) n=159	33,0%	4,7	81,0%
Textlänge 3 (mehr als 20 Zeilen) n=61	37,2%	2,9	73,4%

Tabelle 1: Übersicht der durchschnittlichen Vergabe von Deskriptoren nach Textlänge Testlauf I (n=271)

Ein ähnliches Bild zeigt sich bei der *Klassifikation* (siehe Tab. 2). Die durchschnittliche Anzahl der Klassifikationen, die von MindServer vergeben wird, nähert sich dem Durchschnittswert bei der intellektuellen Erschließung kontinuierlich an. Ebenso nimmt die Überschneidungsmenge durchweg zu. Es wird bei längeren Texten sowohl die Hauptklassifikation häufiger gefunden als auch der Anteil der Dokumente mit einer fehlerhaften Klassifikation nimmt ab.

	Ø-Klass. (MS)	**Ø-Überschneidungen**	**Ø-Überschneidungsmenge an der Ergebnis-menge MS (*Precision*)**	**Ø-Überschneidungen an der Ergebnismenge (intell.)[60] (*Recall*)**	**Hauptklass. gef.**	**Ø-„falscher" Klass. (%)[61]**
Gesamt n=262	3,9	1,4	35,9%	60,9%	72,0%	1,3 (60,4)
Textl. 1 (0-10 Zeilen) n=49	4,5	1,2	26,7%	54,5%	68,0%	1,6 (64,2)
Textl. 2 (10-20 Zeilen) n=155	3,8	1,4	36,8%	58,3%	71,8%	1,4 (63,2)
Textl. 3 (mehr als 20 Zeilen) n=58	3,5	1,5	42,9%	62,5%	75,0%	0,9 (50,0)

Tabelle 2: Übersicht der Zuordnung von Klassifikationen nach Textlänge Testlauf I (n=262)

Die Differenzierung nach den unterschiedlichen *Abstract-Arten* lässt keine signifikanten Unterschiede zwischen den Indexierungsergebnissen des MindServer erkennen. Dies gilt sowohl für die Vergabe der Deskriptoren (vgl. Abb. 10) als auch für die Klassifikation (vgl. Abb. 11).

[60] Die *Recall*-Werte wurden jeweils aus der durchschnittlichen Anzahl intellektuell vergebener Klassifikationen je nach Textlänge errechnet. Diese bewegte sich zwischen 2,2 und 2,4 vergebenen Klassifikationen. Da davon ausgegangen wird, dass die Länge des Abstracts bzw. Autorenreferats keine abhängige Variable für die intellektuelle Indexierung darstellt, wird sie nicht in die Tabelle mit aufgenommen.

[61] Die Angabe in Prozent bezieht sich auf den prozentualen Anteil an der Gesamtmenge.

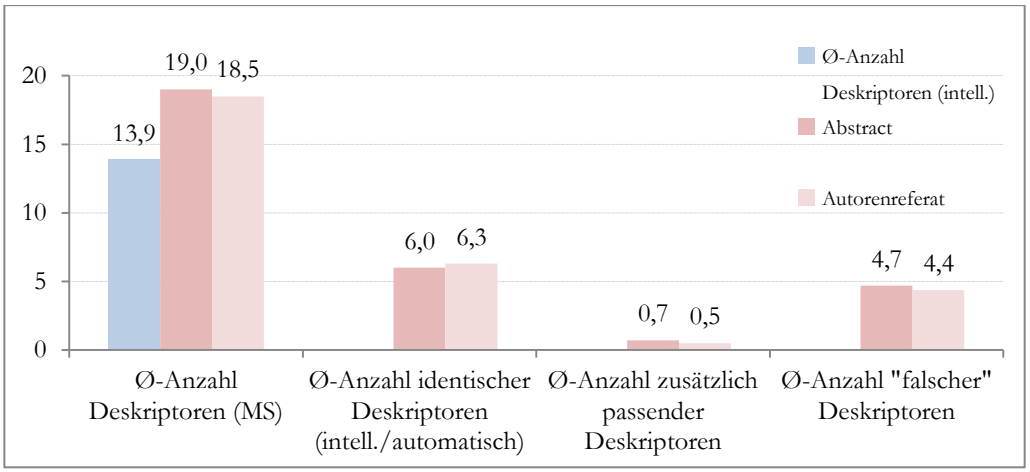


Abbildung 10: Indexierungsunterschiede nach Abstract-Art Testlauf I (Abstract n=167, Autorenreferat n=105)

Die Durchschnittswerte für Dokumente, zu denen ein Autorenreferat vorliegt, liegen bei zentralen Werten näher an der intellektuellen Erschließung. So liegt zum einen bei den Deskriptoren die Überschneidungsmenge geringfügig höher und die Anzahl fehlerhafter Deskriptorenvorschläge minimal niedriger. Zum anderen wird die intellektuell vergebene Hauptklassifikation im Durchschnitt höher gerankt. Gleichwohl wird die Hauptklassifikation bei Dokumenten, zu denen ein Abstract vorliegt, von der Indexierungssoftware insgesamt deutlich häufiger vergeben.

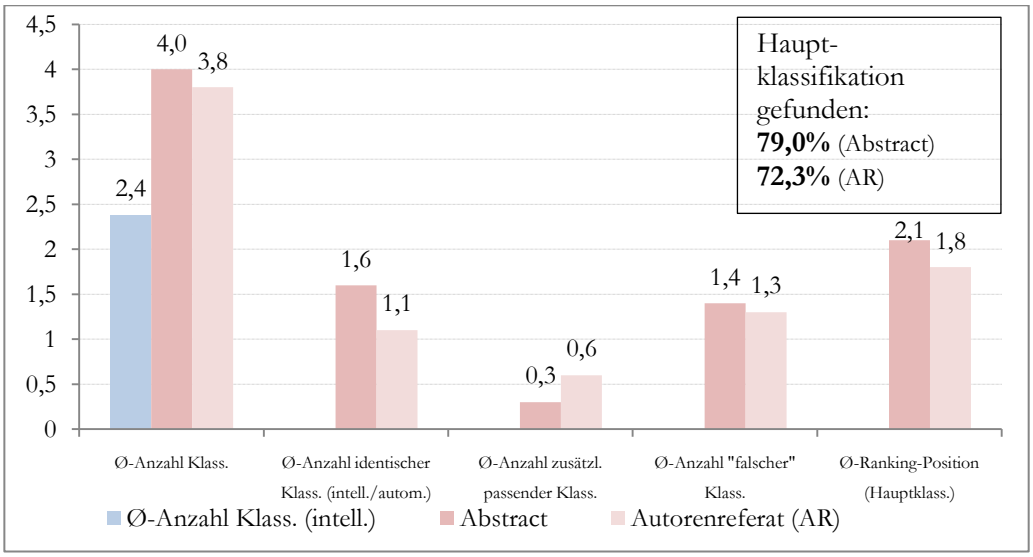


Abbildung 11: Vergleich der Vergabe von Klassifikationen nach Abstract-Art Testlauf I (Abstract n=167, Autorenreferat n=105)

Im Verlauf der Auswertung wurden verschiedene Zwischenschritte bei der Analyse vorgenommen. Sie beeinflussten sehr stark das gesamte weitere Vorgehen für die Testläufe II bis IV. Im Wesentlichen ging es zum einen da-

rum, ein genaueres Bild über die in der Stichprobe enthaltenen Klassifikationen zu erhalten. Es wurde eine Unterteilung der Indexierungsergebnisse des MindServer entsprechend der jeweils intellektuell vergebenen Hauptklassifikation vorgenommen. Zum anderen wurde versucht, ein differenzierteres Bild über die Indexierungsqualität des ersten Testlaufs zu erzielen. Hierfür wurden zusätzliche Variablen, wie etwa verschiedene *cut-off* Level eingeführt, die anhand von Unterstichproben untersucht wurden (siehe hierzu die Übersicht *Zwischenschritte der Analyse* im Anhang).

Zentral für das weitere Vorgehen war die Erstellung einer Übersicht, aus der sowohl die genaue Zusammensetzung der Gesamtstichprobe als auch die Indexierungsergebnisse für die verschiedenen Fachteilgebiete hervorgingen. Differenziert nach den unterschiedlichen Überschneidungsmengen, wie sie aus dem ersten Testlauf hervorgegangen waren, und der jeweils intellektuell vergebenen Hauptklassifikation wurde die Gesamtstichprobe von 271 Dokumenten genau vermessen.[62] Insgesamt wurden den 271 Dokumenten intellektuell 60 verschiedene Hauptklassifikationen zugeordnet. Es zeigte sich, dass die unterschiedliche Gewichtung der einzelnen Unterdisziplinen in Teilen der Zusammensetzung der Datenbank SOLIS entsprach (siehe Kreisdiagramme im Anhang). Es überwiegen die Fachteilgebiete Politikwissenschaft (68 Dokumente bzw. 25 Prozent der Stichprobe) und Soziologie (56 Dokumente bzw. knapp 21 Prozent der Stichprobe), gefolgt von den Bereichen Kommunikationswissenschaften (52 Dokumente bzw. 19 Prozent) und Geisteswissenschaften (41 Dokumente bzw. 15 Prozent).[63]

Aus dem ersten Testlauf geht hervor, dass die höchsten Überschneidungsmengen im Durchschnitt in den Fachteilgebieten Politikwissenschaft (Klassi-

[62] Die Überschneidungsmengen reichten von „keiner“ bis hin zu „mehr als 15 Überschneidungen“ pro Dokumenteinheit. Die Verteilung der Überschneidungsmengen ähnelte dabei einer Normalverteilung.

[63] Das Fachteilgebiet Geisteswissenschaften wurde noch einmal differenziert betrachtet. Ohne die Klassifikation „Sozialgeschichte“, die sich zu einem wichtigen Teilbereich des Fachteilgebiets Soziologie zählen lässt, beträgt die Überschneidungsmenge im Durchschnitt 5,3 Deskriptoren.
Zu einem deutlich geringeren Anteil in der Stichprobe enthalten sind Dokumente zu interdisziplinären und angewandten Gebieten der Sozialwissenschaften (17 Dokumente bzw. 6%), der Demographie (14 Dokumente bzw. 5%) und Erziehungswissenschaft (14 Dokumente bzw. 5%). Am geringsten sind die Bereiche Grundlagen der Sozialwissenschaften, Ethnologie, Psychologie, Wirtschaftswissenschaften, Sozialpolitik sowie Rechts- und Verwaltungswissenschaften mit jeweils unter 2% in der Stichprobe repräsentiert.

fikationsgruppe 10201-10299) mit 7,2 Deskriptoren und Soziologie (Klassifikationsgruppe 10501-10599) mit 6,2 Deskriptoren liegen. Es folgen die Bereiche Geisteswissenschaften (5,8 Deskriptoren) sowie Kommunikationswissenschaften (5,3 Deskriptoren).[64] Ebenso zeigen sich bei der Vergabe der Hauptklassifikation für die verschiedenen Fachteilgebiete Unterschiede bei der erzielten Indexierungsqualität. Während in der Politikwissenschaft in 91,2 Prozent der Dokumente die entsprechende Hauptklassifikation durch die Indexierungssoftware generiert wird, ist dies für das Fachteilgebiet Soziologie in 76,3 Prozent der Dokumente der Fall – gefolgt von Kommunikationswissenschaften (73,1 Prozent), Demographie (71,4 Prozent), Erziehungswissenschaften (64,3 Prozent), Geisteswissenschaften (58,5 Prozent) sowie interdisziplinäre und angewandte Gebiete der Sozialwissenschaften (41,2 Prozent).

Die Unterschiede zwischen den einzelnen Fachteilgebieten bei der automatisch erzielten Indexierungsqualität – einzig gemessen an der intellektuellen Erschließung – sorgten für den Aufbau fachteilgebietsspezifischer Versionen des MindServer für die Testläufe III und IV (siehe hierzu Kap. 3.3.2). Die besseren Indexierungsergebnisse für die Kerngebiete Soziologie und Politikwissenschaft wurden auf den deutlich höheren Anteil an Dokumenten zu diesen Bereichen in der Trainingsmenge zurückgeführt. Daneben wurde angenommen, dass sich die Erkennung von Ähnlichkeiten zwischen den Dokumenten durch die dadurch erzielte deutlichere Homogenität der Trainingsmengen und die Eingrenzung der thematischen und disziplinären Streuung in der Datenkollektion weiter verbessern ließe.

Testlauf II

Ziel des zweiten Testlaufs war es, die Präzision der automatischen Indexierung zu erhöhen. Hierzu wurden einige Veränderungen in den Einstellungen der Indexierungssoftware vorgenommen. Zentrale Veränderungen betrafen die Modifikation des Verhältnisses zwischen Recall und Precision zugunsten

[64] Für die übrigen Bereiche sind die Werte nicht aussagekräftig, da zu wenige Dokumente mit der entsprechenden Hauptklassifikation in der Stichprobe vorhanden waren. Im Einzelnen beträgt der Durchschnittswert für den Bereich Grundlagen der Sozialwissenschaften 7, für Wirtschaftswissenschaften 6,4, für Demografie 5,9, für Erziehungswissenschaften 5,3, für interdisziplinäre und angewandte Gebiete der Sozialwissenschaften 5,2, für Sozialpolitik 5, für Ethnologie 3, für Rechts- und Verwaltungswissenschaften 2,5 und für Psychologie 0,3 Deskriptoren.

der Precision sowie die Einführung unterschiedlicher cut-off Levels für die Vergabe der Deskriptoren und die Zuordnung der Klassifikation(en). (Siehe hierzu erneut Tab. 11 im Anhang).[65]

Im Einzelnen wurden bezogen auf die Vergabe der Deskriptoren die minimale Anzahl der Trainingsdokumente pro Kategorie auf 20 heraufgesetzt. Die maximale Anzahl der Trainingsdokumente, die einem Deskriptor zu seiner Vergabe zugeordnet worden sein durften, blieb unverändert bei 25.000. Das Verhältnis zwischen Recall und Precision wurde zugunsten der Precision von 20,0 auf -20,0 gesetzt. Die Precision wurde dadurch im zweiten Testlauf um 20 Prozent stärker gewichtet als der Recall. Die minimale Textlänge betrug nun, ausgenommen der Stoppwörter, 25 Wörter. Hier war im ersten Testlauf keinerlei Einschränkung vorgenommen worden. Weiterhin wurde keine Einschränkung bei der Mindestanzahl an vergebenen Deskriptoren vorgenommen. Allerdings wurde die maximale Anzahl auf zehn beschränkt.

Bei der *Klassifikation* betrugen minimale und maximale Anzahl der Trainingsdokumente zur Vergabe einer Klassifikation nun ebenfalls 20 und 25.000. Das Verhältnis zwischen *Recall* und *Precision* betrug hier unverändert 20,0. Die minimale Textlänge betrug nun ebenfalls 25 Wörter. Eine minimale Anzahl an vergebenen Klassifikationen wurde – wie bereits beim ersten Testlauf – nicht festgesetzt. Die maximale Anzahl an Klassifikationen wurde nun auf fünf beschränkt. Um den Einfluss der veränderten Systemeinstellungen zu verdeutlichen, werden die Indexierungsergebnisse in den Diagrammen mit den Ergebnissen des ersten Testlaufs in Beziehung gesetzt.[66]

Ergebnisse des zweiten Testlaufs

Für die Vergabe der *Deskriptoren* lässt sich durch die Veränderung des Verhältnisses zwischen *Recall* und *Precision* sowie die Einführung eines *cut-off* Levels eine signifikante Erhöhung der *Precision* erzielen (siehe Abb. 12). Diese steigt von 32,4 auf 53,0 Prozent, was einem Anstieg um 65,6 Prozent entspricht. Wie zu erwarten war, geht diese Erhöhung mit einer Absenkung

[65] Aus technischen Gründen mussten für diesen Testlauf zunächst sämtliche Dokumente der Gesamtstichprobe in ein XML-Format transformiert werden.
[66] Zum Teil ließ sich dies nur in unübersichtlicher Weise graphisch umsetzen. In diesen Fällen wurde auf den Vergleich in den Diagrammen verzichtet.

des *Recall* einher. So sinkt der *Recall* von 44,0 auf 31,0 Prozent, was einer Absenkung um 29,6 Prozent entspricht.

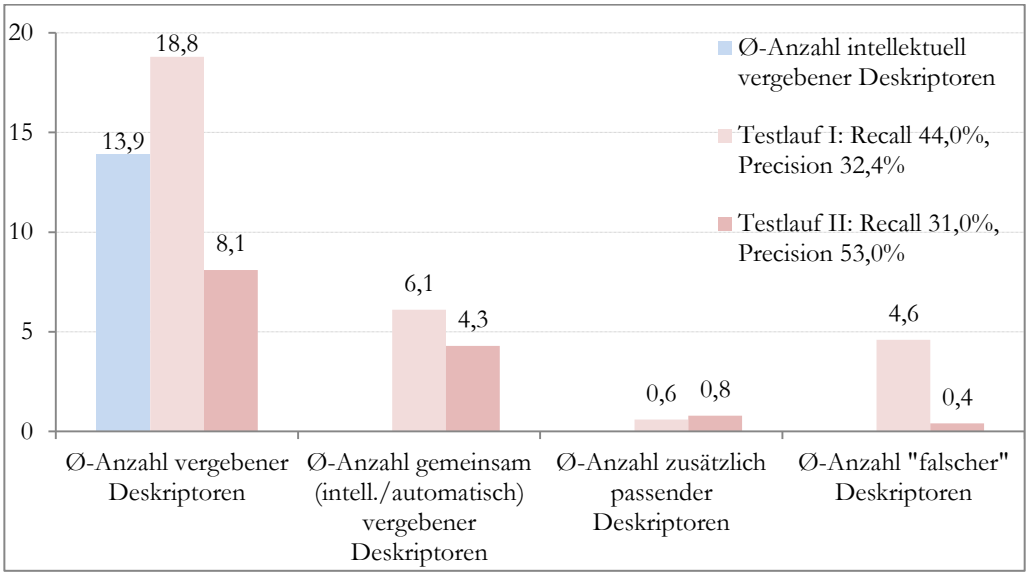


Abbildung 12: Vergleich der Vergabe von Deskriptoren Testläufe I (n=271) und II (n=263)

Als ein weiteres positives Ergebnis geht aus dem Vergleich der beiden Testläufe hervor, dass sich die durchschnittliche Anzahl fehlerhafter Deskriptorzuordnungen sehr deutlich verringert. Da die Auswertung der ersten beiden Testläufe zu weiten Teilen von unterschiedlichen Indexierern vorgenommen wurde, lässt sich dies partiell vermutlich auch auf die Inter-Indexiererkonsistenz zurückführen.

In einem Vergleich der Gesamtverteilungen von *Recall* und *Precision* auf die Stichprobe zeigen sich die Verschiebungen in die unterschiedlichen Richtungen sehr anschaulich (siehe Abb. 13 und 14). Deutlich zentrieren sich die *Recall*-Werte im zweiten Testlauf auf 20 bis 39 Prozent.

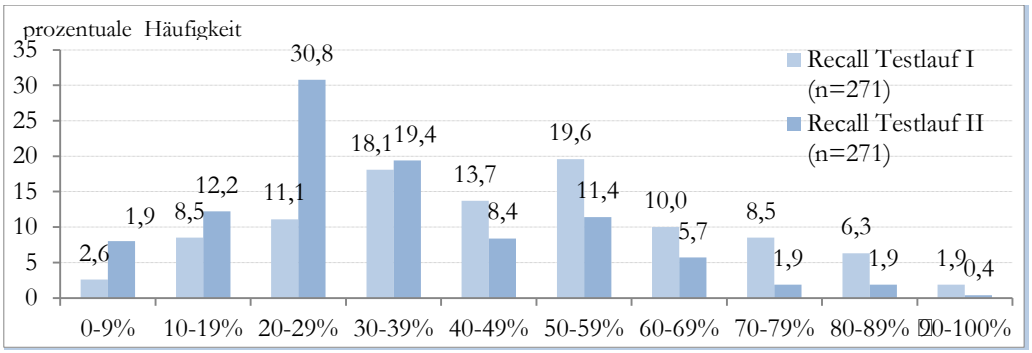


Abbildung 13: Verteilung der *Recall*-Werte auf die Gesamtstichprobe Testläufe I (n=271) und II (n=261)[67]

Breiter angelegt erscheint die Verschiebung der *Precision*-Werte. Während die meisten automatisch generierten Indexierungsergebnisse im ersten Testlauf einen *Precision*-Wert zwischen 20 und 59 Prozent erzielen, verteilen sich die Werte im zweiten Testlauf zwischen 30 und 79 Prozent.

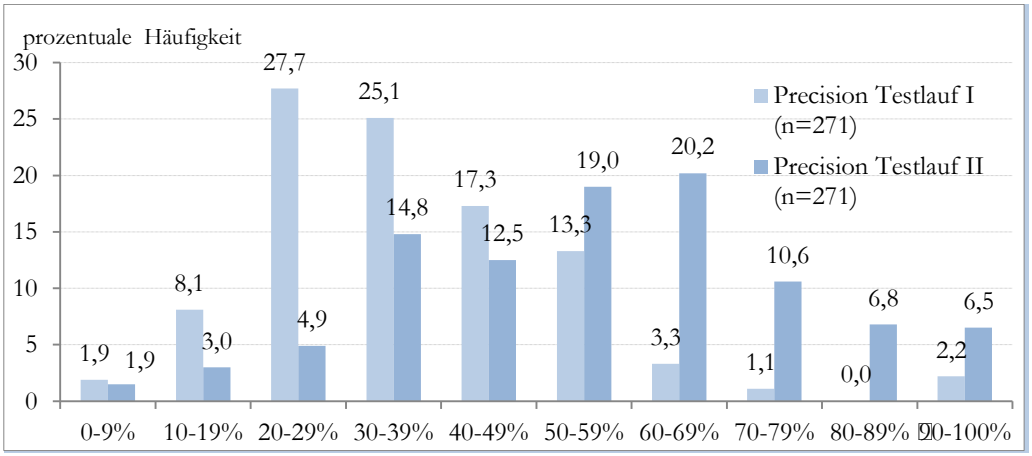


Abbildung 14: Verteilung der *Precision*-Werte auf die Gesamtstichprobe Testläufe I (n=271) und II (n=261)[68]

Mit Blick auf die verschiedenen Dokumentarten wird deutlich, dass sich die Unterschiede zwischen den Dokumentarten nivellieren (siehe Abb. 15).

[67] Die Anzahl der Dokumente, die beim zweiten Testlauf von der Indexierungssoftware erschlossen wurde, betrug 263. Allerdings ließen sich die Daten im Textprogramm nicht mehr bearbeiten.

[68] Auch in dieser Abbildung ließen sich die Daten im Textprogramm nicht mehr nachträglich ändern.

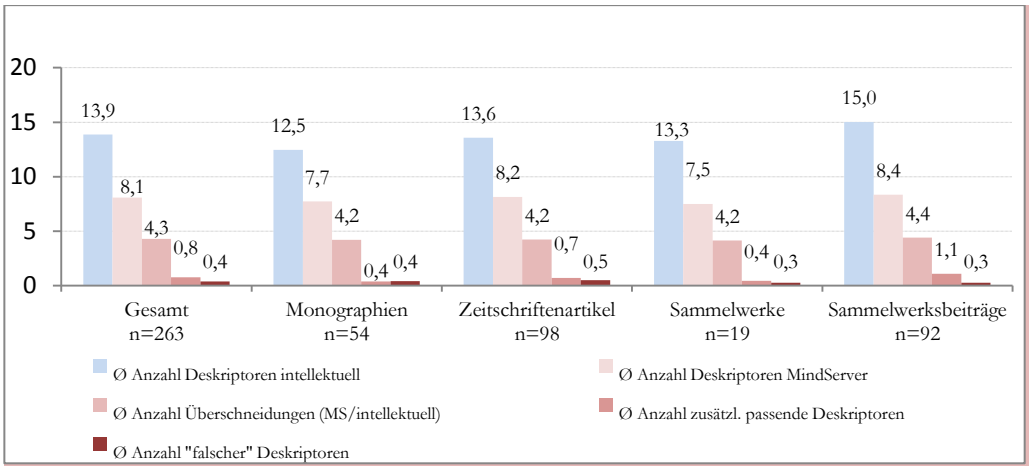


Abbildung 15: Vergabe der Deskriptoren nach Dokumentarten Testlauf II (n=263)[69]

Sowohl die Anzahl der vergebenen Deskriptoren als auch die Menge von Deskriptoren, die intellektuell sowie maschinell vergeben wird, nähern sich zwischen den Dokumentarten einander an. Das Häufigkeitsverhältnis bei der Vergabe der Deskriptoren zwischen den Dokumentarten entspricht dem ersten Testlauf. Sammelwerksbeiträge generieren die meisten, Gesamtaufnahmen von Sammelwerken hingegen die geringste Anzahl an Deskriptoren. Analog zum Gesamtergebnis für den zweiten Testlauf (siehe Abb. 12) sinkt die durchschnittliche Anzahl fehlerhafter Zuweisungen von Deskriptoren deutlich.

In Bezug auf die Vergabe der *Klassifikation* zeigt sich im Vergleich zum ersten Testlauf für einige zentrale Auswertungskategorien eine leichte Verbesserung der Indexierungsqualität (siehe Abb. 16). Die Einführung eines *cut-off* Levels von fünf führt zu einer Reduktion der durchschnittlichen Anzahl automatisch generierter Klassifikationen, hierauf wird ebenso zurückgeführt, dass auch die Anzahl fehlerhafter Klassifikationszuweisungen im Durchschnitt deutlich zurückgeht. Dies führt auch hier zu einer geringfügigen Erhöhung der *Precision* auf 40,0 Prozent (35,9 Prozent) auf der einen und einem leichten Absinken des *Recall* auf 58,3 Prozent (60,9 Prozent) auf der anderen Seite. Dadurch, dass das Verhältnis zwischen *Recall* und *Precision* bei den Einstellungen nicht verändert wurde, bleibt die durchschnittliche Anzahl der

[69] Die weitere Reduktion der Dokumentenzahl, für die ein automatisch generiertes Indexierungsergebnis vorlag, ist in erster Linie auf die Festlegung der Testlänge auf 25 Wörter zurückzuführen.

Überschneidungsmenge nahezu identisch.[70] Da die Systemeinstellungen für die Klassifikation zwischen erstem und zweitem Testlauf bis auf die Einführung des *cut-off* Levels identisch waren, lässt sich die deutliche Dynamik beim Ranking der zugewiesenen Klassifikationsnotationen einzig hierauf zurückführen.

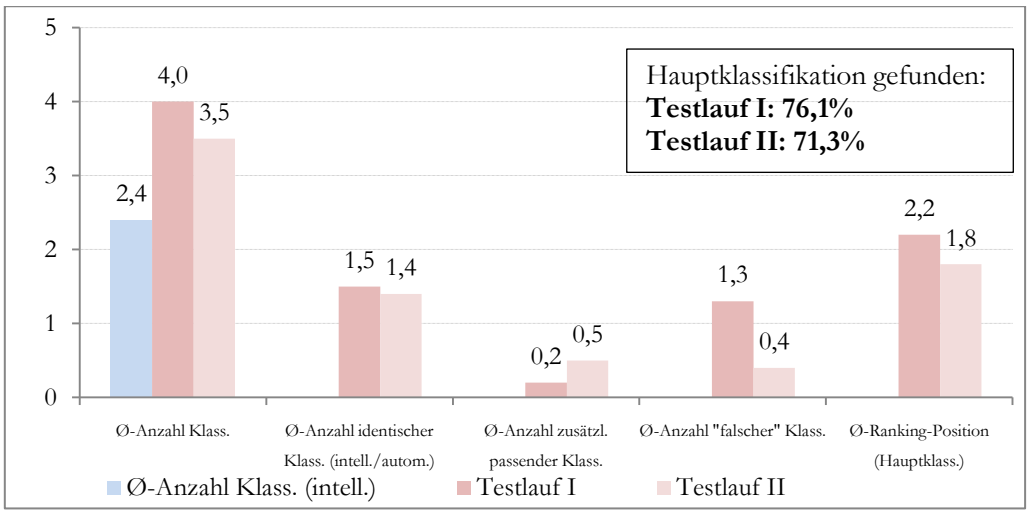


Abbildung 16: Vergleich der Vergabe von Klassifikationen Testläufe I (n=262) und II (n=254)

Eine Nivellierung der Indexierungsunterschiede zwischen den *Dokumentarten* beim zweiten Testlauf geht auch aus den Klassifikationsergebnissen hervor. Dabei zeigen sich zwischen den Dokumentarten dieselben unterschiedlichen Häufigkeiten bei der Klassifikationszuweisung wie beim ersten Testlauf. Auffällig ist auch hier die deutliche Abnahme fehlerhafter Klassifikationszuordnungen.

[70] In ähnlicher Weise wie für die Auswertung der Deskriptorenvergabe lassen sich diese Werte in Bezug auf Inter- sowie Intra-Indexiererkonsistenz diskutieren. Vor dem Hintergrund, dass das Verhältnis zwischen *Recall* und *Precision* bei den Systemeinstellungen zwischen erstem und zweitem Testlauf identisch blieb, gilt dies vor allem für die Zunahme der durchschnittlichen Anzahl zusätzlich passender Klassifikationen.

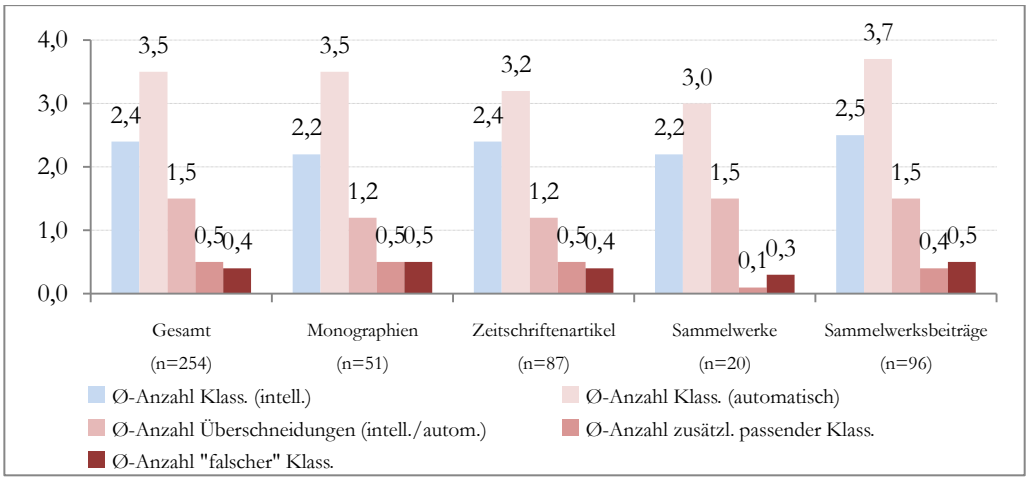


Abbildung 17: Vergabe der Klassifikation nach Dokumentart Testlauf II (n=254)

Bei einem Vergleich der Indexierungsergebnisse in Bezug auf die *Abstract-Art* zeigen sich exakt dieselben minimalen Unterschiede wie bereits beim ersten Testlauf (siehe Abb. 11). Die Indexierungsergebnisse zu Dokumenten, die anhand eines Autorenreferats erschlossen wurden, bewegen sich bei der Vergabe der Deskriptoren geringfügig näher an der intellektuellen Indexierung.

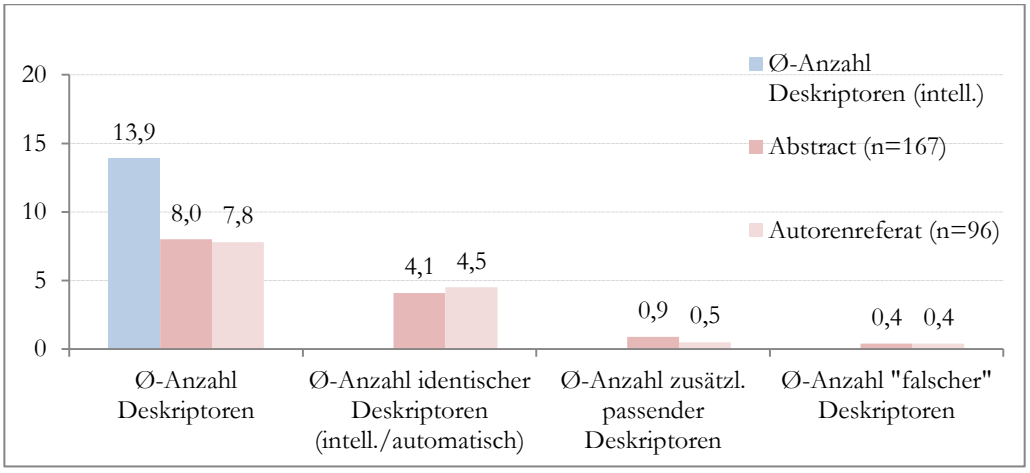


Abbildung 18: Vergleich der Vergabe von Deskriptoren nach Abstract-Art Testlauf II (n=263)

Bei der Klassifikation ist die Indexierungssoftware hingegen bei denjenigen Dokumenten näher an der intellektuellen Erschließung, die mit einem Abstract erschlossen wurden (siehe Abb. 19). Allerdings zeichnen sich hier die Indexierungsergebnisse zu Dokumenten, die ein Autorenreferat aufweisen, durch mehr zusätzlich passende Klassifikationen sowie eine höhere durchschnittliche Ranking-Position der Hauptklassifikation aus. Wie bereits für die Gesamtstichprobe zeigt sich auch hier, dass die Abstract-Art keinen signifikanten Einfluss auf das Indexierungsergebnis hat.

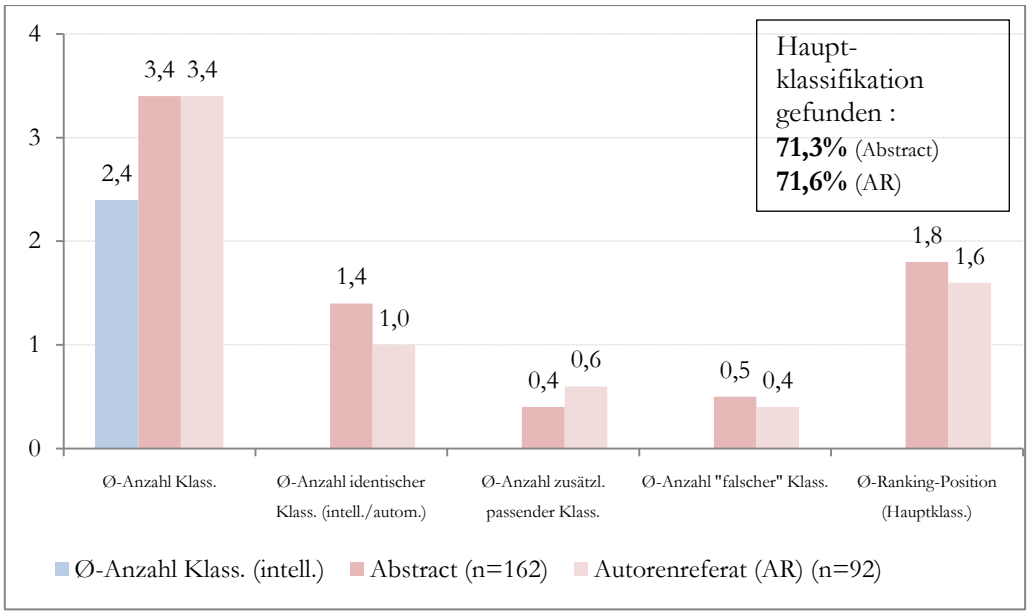


Abbildung 19: Vergabe der Klassifikation nach Abstract-Art Testlauf II (n=254)

Zusammenfassend stellen für die Indexierungssoftware sowohl Dokument- als auch Abstract-Art keine entscheidenden Variablen dar, die sich auf die Erschließungsqualität – gemessen an der intellektuellen Indexierung – auswirken.

Im Anschluss an diesen Teil der Auswertung des zweiten Testlaufs wurde erneut eine Differenzierung der Indexierungsergebnisse nach Fachteilgebieten vorgenommen. Ein weiteres Mal zeigt sich, dass in den Kernbereichen der Datenbank SOLIS sowohl bezogen auf die Vergabe der Deskriptoren als auch hinsichtlich der Klassifikation durchschnittlich höhere Überschneidungsmengen erzielt werden als in den Randbereichen. Während bei den Dokumenten aus den Fachteilgebieten Soziologie und Politikwissenschaft im Durchschnitt 4,9 bzw. 5,0 der im zweiten Testlauf insgesamt durchschnittlich acht Deskriptoren sowohl intellektuell als auch maschinell vergeben werden, liegt die Überschneidungsmenge bei den Dokumenten der Kommunikationswissenschaften beispielsweise bei lediglich 3,7 Deskriptoren. Ebenso erzielt die Indexierungssoftware bei der Vergabe der Klassifikationen für die Kernbereiche der Datenbank ein besseres Ergebnis als für die Randbereiche. So wird bei den Kerngebieten Soziologie und Politikwissenschaft die Hauptklassifikation in 75,5 Prozent bzw. 90,6 Prozent der Fälle ebenso maschinell generiert,

während der Anteil bei der Demografie etwa bei lediglich 53,9 Prozentpunkten liegt.[71]

Zwischenergebnisse

Ebene der Dokumentenkollektion

Aus dem Vergleich der beiden Testläufe auf der Ebene der Dokumentenkollektion geht für die Vergabe der *Deskriptoren* hervor, dass mit Einführung eines *cut-off* Levels und einer Veränderung des Verhältnisses zwischen *Recall* und *Precision* zugunsten der *Precision* sich die Präzision der automatischen Indexierungsergebnisse deutlich erhöhen lässt.

	Testlauf I (n=271)	**Testlauf II** (n=263)
Recall	44,0% (*48,6%*)[72]	30,9% (*36,4%*)
Precision	32,4% (*35,8%*)	53,0% (*62,3%*)
Ind.-Konsistenz	37,3% *(41,3%)*	39,0% *(46,0%)*

Tabelle 3: *Recall*- und *Precision*-Werte sowie Indexierungskonsistenz bei der Vergabe von Deskriptoren für die Testläufe I und II

Auf diese Veränderungen der Systemeinstellungen wird ebenso zurückgeführt, dass sich die Anzahl fehlerhafter Deskriptorvorschläge sehr stark reduziert. Die Indexierungskonsistenz erhöht sich hingegen nur geringfügig. Sie liegt bei knapp 40 Prozent. Das Gesamtverhältnis zwischen identischen sowie insgesamt – sowohl maschinell als auch intellektuell vergebenen Deskriptoren – verändert sich nur kaum.

Weniger eindeutig erscheinen die Ergebnisse für die maschinelle Zuweisung der *Klassifikation(en)*. Einerseits zeigt sich in Verbindung mit der Einführung eines *cut-off* Levels auch hier eine leichte Erhöhung des *Precision*-Wertes. Andererseits sorgt diese dafür, dass die intellektuell vergebene Hauptklassifikation seltener von der Indexierungssoftware vorgeschlagen wird.[73] Tendenziell wird die Hauptklassifikation beim zweiten Testlauf im Durchschnitt hö-

[71] Vgl. exemplarisch die Gegenüberstellung von Kern- und Randbereichen an den Beispielen Soziologie, Politik- und Erziehungswissenschaft im Anhang Tab. 10.
[72] Diese *Recall*- und *Precision*-Werte in Klammern würden bei einer Einberechnung der durchschnittlichen Anzahl zusätzlich passender Deskriptoren erzielt.
[73] Im ersten Testlauf wurde die intellektuell vergebene Hauptklassifikation in 76,1 Prozent der Fälle auch von der Indexierungssoftware generiert. Im zweiten Testlauf betrug der Anteil lediglich 71,3 Prozent.

her gerankt als beim ersten Testlauf. Die Indexierungskonsistenz erhöht sich auch hier nur geringfügig.

	Testlauf I (n=262)	**Testlauf II** (n=254)
Recall	63,0% (70*,8%)*	58,3% (*79,2%)*
Precision	37,5% (*42,5%)*	40,0% (*54,3%)*
Ind.-Konsistenz	45,2% *(52,0%)*	46,4% *(61,8%)*

Tabelle 4: Recall- und Precision-Werte sowie Indexierungskonsistenz für die Zuordnung von Klassifikationsnotationen für die Testläufe I und II

Daneben geht aus dem ersten Testlauf hervor, dass die automatisch generierten Erschließungsergebnisse mit zunehmender Länge von Abstract und Autorenreferat näher an der intellektuellen Indexierung liegen. Dokument- und Abstract-Art haben hingegen bei beiden Testläufen keinen signifikanten Einfluss auf die Erschließungsqualität der Indexierungssoftware.

Im Rahmen einer stichprobenartigen Untersuchung geht daneben in beiden Testläufen aus dem Vergleich der Erschließungsergebnisse für die verschiedenen Fachteilgebiete hervor, dass die Indexierungssoftware – gemessen an der intellektuellen Erschließung – tendenziell bessere Ergebnisse bei den Kernbereichen der Datenbank SOLIS als bei den Randbereichen erzielt. Dies gilt sowohl für die Vergabe der Deskriptoren als auch eingeschränkt für die Zuweisung der Klassifikationsnotationen.

	Deskriptorenvergabe		**Klass.zuweisung**	
	Testlauf I	Testlauf II	Testlauf I	Testlauf II
Soziologie (n=56)				
Recall %	46,6 *(51,8)*	32,5 *(37,8)*	56,5 *(65,2)*	56,5 *(82,6)*
Precision %	31,5 *(35,1)*	51,4 *(62,7)*	30,2 *(34,9)*	36,1 *(52,8)*
Hauptklass. gef. %			78,2	78,4
Ranking-Position			1,7	1,6
Politik (n=68)				
Recall %	46,0 *(48,3)*	32,2 *(37,2)*	82,6 *(87,0)*	73,9 *(82,6)*
Precision %	36,1 *(37,9)*	60,1 *(69,6)*	47,5 *(50,0)*	47,2 *(52,8)*
Hauptklass. gef. %			92,4	87,3
Ranking-Position			1,7	1,5
Pädagogik (n=14)				
Recall %	35,9 *(36,0)*	18,0 *(18,2)*	47,4 *(57,9)*	42,1 *(57,9)*
Precision %	44,7[74] *(44,8)*	43,0 *(43,5)*	47,4 *(57,9)*	30,8[75](42,0)

[74] Aufgrund der relativ niedrigen durchschnittlichen Anzahl automatisch generierter Deskriptoren (11,4) – was der Annahme einer schwierigen Kontext-Erkennung bei den Randbereichen der Datenbank entspricht – ist der *Precision*-Wert verhältnismäßig hoch.

[75] Die Einführung des *cut-off* Levels für die Klassifikation im zweiten Testlauf sorgte für einen deutlichen Anstieg der durchschnittlichen Klassifikationsvergabe im Vergleich zum

Hauptklass. gef. %	80,0	63,6
Ranking-Position	1,5	1,6

Tabelle 5: Indexierungsergebnisse für die Fachteilgebiete Testlauf I und II

In den nachfolgenden beiden Testläufen wurde an dieses Ergebnis einer unterschiedlichen Erschließungsqualität zwischen den Kern- und Randbereichen der Datenbank SOLIS angeknüpft.

Ebene der Einzeldokumente

Auf der Ebene der einzelnen Dokumente zeigt sich für die Vergabe der Deskriptoren, dass die Einstellungsveränderungen zu einer Neugewichtung der inhaltlichen Aspekte einer Publikation führen können. Mit Erhöhung der *Precision* findet die Ähnlichkeitserkennung zu Dokumenten der Trainingsmenge in einem enger gefassten Bezugsraum statt. Die Vergabe der Deskriptoren durch die Indexierungssoftware scheint strenger abzulaufen. Dies wirkt sich zum einen positiv aus, indem beim zweiten Testlauf neue zusätzliche Deskriptoren generiert werden, die ebenso intellektuell vergeben worden waren. Dazu gehören sowohl sachliche Schlagwörter, für deren Vergabe mitunter eine Kompositazerlegung und Kontexterkennung notwendig waren, als auch Geographika.[76] Zum anderen kann die Erhöhung der *Precision* durch den enger gefassten Abgleich mit den Dokumenten aus der Trainingsmenge jedoch auch zu einer starken Neugewichtung der inhaltlichen Aspekte führen: Hauptaspekten kann eine deutliche geringere, Nebenaspekten hingegen eine gestiegene Relevanz zugesprochen werden.[77]

Daneben kann die Veränderung der Systemeinstellungen auch dazu führen, dass von der Indexierungssoftware gar keine Erschließung vorgenommen

---

ersten Testlauf. Damit verbunden reduzierte sich die *Precision* der Indexierungsergebnisse deutlich.

[76] Ein Beispiel bildet die Zuordnung des Deskriptors „palästinensisch-israelischer Konflikt" zu einer Publikation über die politische Ausrichtung russischer Neueinwanderer in Israel. Dieser Deskriptor wird im zweiten Testlauf zusätzlich vergeben, da in der Trainingsmenge eine Vielzahl von Publikationen vorhanden sind, die im Kontext der Themenbereiche „Israel" und „politische Einstellung" diesen Deskriptor intellektuell zugeordnet bekommen haben.

[77] Ein anschauliches Beispiel liefert ein Vergleich der Indexierungsergebnisse zu einer Publikation zum Polenbild der deutschen Rechten. Im Abstract wird erwähnt, dass der Polendiskurs unter anderem von Polens Rolle im zweiten Irak-Krieg geprägt war. Während beim ersten Testlauf der Konfidenzwert für den Deskriptor „Polen" 0,44 beträgt, hingegen das Geographikum „Irak" nicht vergeben wird, dreht sich das Verhältnis beider Deskriptoren beim zweiten Testlauf um. Der Konfidenzwert für den Deskriptor „Irak" beträgt 0,36, während „Polen" gar nicht verzeichnet wird, obgleich die Publikation ausschließlich vom Bild Polens in Deutschland handelt.

wird, obgleich die formalen Voraussetzungen dafür erfüllt werden. Dies betrifft vor allem Gesamtaufnahmen von Sammelwerken. Die Vielzahl unterschiedlicher Aspekte, die häufig sowohl aus dem einleitenden Autorenreferat als auch aus dem angefügten Inhaltsverzeichnis hervorgehen kann, führt mitunter dazu, dass keinerlei Kontexterkennung bzw. Ähnlichkeitsabgleich mit anderen Dokumenten möglich ist.[78] Doch auch Abstracts und Autorenreferate zu Artikeln in Zeitschriften und Beiträgen in Sammelwerken, die sich durch sehr spezifische Themenstellungen auszeichneten, liefern mitunter keine Indexierungsergebnisse.[79]

Bei der Zuordnung der *Klassifikation* durch die Indexierungssoftware zeigt sich eine Schwierigkeit bei der Zuordnung und Gewichtung der einzelnen Wissenschaftsdisziplinen, die für die Einordnung der Publikationen relevant erscheinen, bzw. der einzelnen inhaltlichen Aspekte, die in der Publikation angesprochen werden. Klassifikationen, die intellektuell auf einem der unteren Ränge als Nebenklassifikationen eingestuft wurden, werden nicht selten von der Indexierungssoftware als Hauptklassifikationen zugeordnet – und umgekehrt.[80] Wird die Anzahl an Klassifikationen, die pro Indexat vergeben werden können, nicht eingeschränkt kann dies zu einer Verzerrung der Klassifikationszuordnung führen. So werden in einigen Fällen, in denen keine eindeutige Zuweisung durch den Abgleich mit anderen Dokumenten erfolgen kann, sehr viele Klassifikationsvorschläge aus einer Klassifikationsgruppe generiert.[81]

---

[78] Dies betraf etwa einen Sammelband mit dem Titel „Gesellschaft und Kultur“, dessen Beiträge unter anderem sowohl auf Sozialmedizin, Sexualität und soziale Beziehungen als auch auf politische und weltanschauliche Phänomene wie Marxismus, Bolschewismus, Krieg und Massenbewegungen eingingen. Aus den Testläufen III und IV, in denen der *Precision*-Wert erneut niedriger lag, gingen für diese Publikationen zum Teil erneut automatisch generierte Indexierungsergebnisse hervor.

[79] Als Beispiel lässt sich ein Aufsatz mit dem Titel „Allyfying Leibniz“ zu Aspekten von Kompossibilität und Diegese anführen. Darin wird unter anderem die von Leibniz entwickelte Weltenpyramide auf die Fernsehserie Ally McBeal bezogen.

[80] Exemplarisch lässt sich eine Publikation anführen, für die in den ersten beiden Testläufen die intellektuell vergebene Hauptklassifikation („Sozialgeschichte, historische Sozialforschung“) zunächst auf dem vierten von fünf Rängen angezeigt wird, die intellektuell generierte Nebenklassifikation („politische Willensbildung, politische Soziologie“) hingegen als Hauptklassifikation gerankt wird. (In den Testläufen III und IV kehrte sich das Verhältnis hingegen um.)

[81] Unter Umständen führte dies im ersten Testlauf dazu, dass aus einem Fachteilgebiet mehr als fünf Klassifikationen vergeben werden. Eine Publikation, die in den Bereich der Kommunikationswissenschaften (10800-10899) fällt, führt etwa zu sieben unterschiedlichen Klassi-

3.3.2 Die Testläufe III und IV

Die Unterschiede bei der automatischen Indexierung zwischen Kern- und Randbereichen der Datenbank SOLIS führten zum Aufbau fachteilgebietsspezifischer Versionen des MindServer.[82] Hiermit war die Hypothese verbunden, dass sich durch die Eingrenzung der Trainingsmenge auf einzelne Fachteilgebiete die Ähnlichkeitserkennung zwischen den Dokumenten und damit die Erschließungsergebnisse weiter verbessern ließen.

Für die Testläufe III und IV wurden hierfür aus der gesamten Dokumentenkollektion der Datenbank SOLIS anhand der intellektuell vergebenen Klassifikationsnotationen die Datenbestände zu drei verschiedenen Fachteilgebieten selektiert.[83] Hierzu gehörten die Fachteilgebiete Soziologie, Politikwissenschaft sowie Geisteswissenschaften. Diese Auswahl diente dazu, exemplarisch die Indexierungsergebnisse des MindSever für Kern- und Randbereiche miteinander zu kontrastieren und vor diesem Hintergrund das Schichtenmodell nach Jürgen Krause (1996), das bereits zu einem sehr frühen Zeitpunkt Eingang in die Beschäftigung mit automatischer Indexierung in der sozialwissenschaftlichen Fachinformation gefunden hat (siehe oben), zu diskutieren.[84]

Für jedes dieser Fachteilgebiete wurde somit eine eigene Version des MindServer aufgebaut, die jeweils ausschließlich auf der Grundlage der entsprechenden selektierten Dokumentenmenge trainiert wurde. Für die Soziologie

---

fikationen in diesem Bereich (10800, 1080401, 1080404, 1080405, 1080407, 1080411, 1080412). Dies erschwert mitunter auch die Bewertung der einzelnen maschinell generieten Klassifikationsvorschläge.

[8282] Dies entsprach eher einem explorativen Vorgehen. Es zeigte sich bereits relativ bald, dass die praktische Umsetzung eines solchen Aufbaus fachteilgebietsspezifischer MindServer-Versionen nur unter sehr großem Aufwand möglich war.

[83] Unter Programmierung einer XSL-Datei sowie eines Java-Programms zur Aktivierung dieser Datei wurde hierzu der gesamte Datenbestand der Datenbank SOLIS durchlaufen. Jene Dokumente, deren intellektuell vergebene Klassifikationen in die Klassifikationsbereiche 10200 bis 10299, 10500 bis 10599 sowie 30000 bis 39900 fielen, wurden herausgefiltert und im XML-Format abgelegt. Eine Filterung des gesamten Dokumentbestands einzig anhand der entsprechenden Hauptklassifikation erschien aus technischen Gründen nicht möglich. Unter Umständen hätte dies allerdings auch die Trainingsmenge auf eine ziemlich kleine Dokumentenzahl reduziert, was sich negativ auf die Indexierungsleistung ausgewirkt hätte.

[84] Die Auswahl des Randbereichs Geisteswissenschaften hing mit der Zusammensetzung der Gesamtstichprobe zusammen. So stellten die Geisteswissenschaften in der Stichprobe einen der größeren Datenbestände zu den Randgebieten dar. Um die Unterschiede zwischen Kern- und Randbereichen zu veranschaulichen, wird in den Diagrammen versucht, die Indexierungsergebnisse für die drei Fachteilgebiete jeweils gemeinsam darzustellen.

umfasste diese circa 142.000 Dokumente, für die Politikwissenschaft knapp 75.000 Dokumente und für die Geisteswissenschaften etwa 54.500 Dokumente. Analog wurden auch aus der Gesamtstichprobe von 271 Dokumenten diejenigen Dokumente selektiert, deren intellektuell vergebene Hauptklassifikation in die entsprechenden Klassifikationsbereiche fiel. Im Einzelnen waren dies 56 Dokumente für das Fachteilgebiet Soziologie, 68 Dokumente für den Bereich Politikwissenschaft und 40 Dokumente für den Bereich Geisteswissenschaften.

Testlauf III

Für den dritten Testlauf wurden die Standardeinstellungen des MindServer übernommen, wie sie bereits für den ersten Testlauf galten (siehe Tab. 11). Während Minimum und Maximum der Trainingsdokumente pro Kategorie somit bei einem bzw. 25.000 Dokumenten lagen, wurden weder die Anzahl der vergebenen Deskriptoren noch die Anzahl der generierten Klassifikationen eingeschränkt. Ebenso wurde auch keine minimale Textlänge festgesetzt, um einen Deskriptor bzw. eine Klassifikation zu vergeben. Das Verhältnis zwischen *Precision* und *Recall* wurde zugunsten des *Recall* bei +20,0 belassen.

Ergebnisse des dritten Testlaufs

Aus dem Vergleich der automatisch generierten Indexierungsergebnisse für die einzelnen Fachteilgebiete geht hervor, dass im Randbereich der Datenbank weniger Deskriptoren vergeben werden (siehe Abb. 20). Dies bestätigt die Annahme, wonach der Indexierungssoftware die Erkennung von Ähnlichkeiten zu anderen Dokumenten des Fachteilgebiets im Randbereich schwerer fällt, da die Trainingsmengen der Teilgebiete im Randbereich geringer ausfallen.[85]

[85] Es wird davon ausgegangen, dass die Größe der Trainingsmenge allein allerdings nicht den Ausschlag gibt für die unterschiedlichen Indexierungsergebnisse. Vielmehr wird angenommen, dass sich ebenso sowohl der Radius eines Fachteilgebiets als auch die Zusammensetzung der Trainingsmenge deutlich auf die automatische Indexierung auswirken. Ein tieferer Einblick in die Zusammensetzung der unterschiedlichen Trainingsmengen ist an dieser Stelle jedoch nicht möglich.

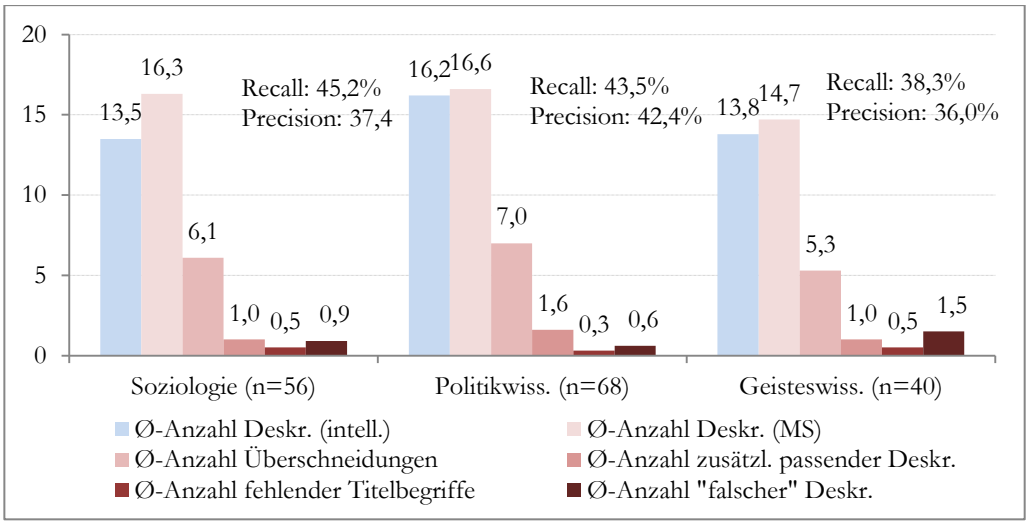


Abbildung 20: Vergleich der Vergabe von Deskriptoren zwischen den Fachteilgebieten Soziologie, Politik- und Geisteswissenschaften Testlauf III

Hierauf werden ebenso die im Vergleich zu den Kernbereichen niedrigere Überschneidungsmenge sowie die höhere Anzahl fehlerhafter Deskriptorzuordnungen zurückgeführt. Dieses Verhalten der Indexierungssoftware wirkt sich auf die Werte für *Recall* und *Precision* aus. Auch nach Aufbau einer teilgebietsspezifischen Trainingsmenge liegt der *Recall* für den Randbereich in ähnlicher Weise wie bereits in den Voruntersuchungen deutlich unterhalb des Durchschnitts für die Gesamtstichprobe (vgl. Abb. 6).

Für die Kernbereiche lässt sich durch den Aufbau der unterschiedlichen Trainingsmengen für die Fachteilgebiete eine deutliche Erhöhung der *Precision* erzielen. Im Vergleich zu den Ergebnissen für die Fachteilgebiete Soziologie und Politikwissenschaft aus dem ersten Testlauf – in beiden Testläufen waren die Einstellungen der Indexierungssoftware identisch – ergibt sich für das Fachteilgebiet Soziologie ein prozentualer Anstieg der *Precision* um knapp 20 Prozent (18,9 Prozent). Für das Fachteilgebiet Politikwissenschaft hingegen lässt sich ein prozentualer Anstieg der *Precision* um knapp 18 Prozent (17,5 Prozent) verzeichnen.[86]

[86] Nachträglich wurden auch die fachteilgebietsspezifischen *Recall*- und *Precision*-Werte des ersten Testlaufs für den Bereich Geisteswissenschaften berechnet. Auch hier zeigt sich, dass durch den Aufbau einer fachspezifischen Trainingsmenge eine Erhöhung der *Precision* erzielt wird. Ihr Anteil beträgt 12,5 Prozent.

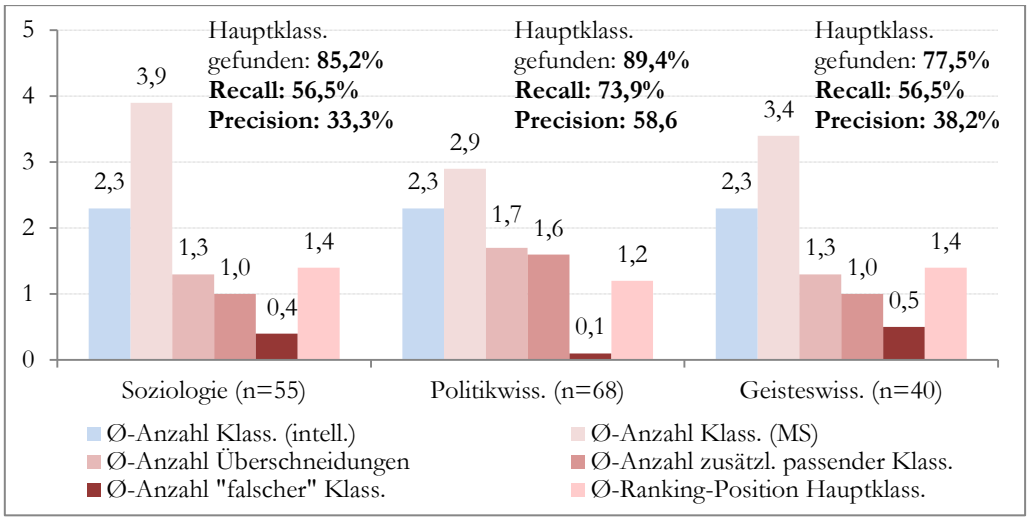


Abbildung 21: Vergleich der Klassifikation zwischen den Fachteilgebieten Soziologie, Politik- und Geisteswissenschaften Testlauf III

Auch für die *Klassifikation* lässt sich durch den Aufbau fachteilgebietsspezifischer Trainingsmengen die *Precision* (leicht) erhöhen. Der Aufbau der fachspezifischen Trainingsmengen führt zu einer Reduktion der Klassifikationszuweisungen.[87] Verbunden mit der deutlich geringeren Anzahl an Kategorien bei der Klassifikation nähern sich hier die Indexierungsergebnisse zwischen Kern- und Randbereichen einander an.[88] Ebenso wird die intellektuell vergebene Hauptklassifikation beim dritten Testlauf im Vergleich zu den vorherigen im Durchschnitt höher gerankt (vgl. Tab. 3). Häufiger als beim ersten Testlauf wird die intellektuell generierte Hauptklassifikation in den Fachteilgebieten Soziologie und Geisteswissenschaften vergeben. Außerdem lässt sich die Anzahl fehlerhafter Klassifikationsvorschläge deutlich reduzieren.[89] Durchweg erzielt die Indexierungssoftware für das Fachteilgebiet Politikwissenschaft – allein gemessen an der intellektuellen Erschließung – die besten Ergebnisse. Nicht nur kommt die Anzahl der automatisch generierten Klassifikationen dem Durchschnittswert der intellektuellen Klassifikation am nächsten, auch die Anzahl der Überschneidungen und zusätzlich passender Klassifikationen ist am höchsten, die Anzahl fehlerhafter Klassifikationen am

[87] Ebenfalls nachträglich wurde die durchschnittliche Anzahl der Klassifikationszuweisungen beim ersten Testlauf für die drei Fachteilgebiete ermittelt. Diese betragen 4,3 (Soziologie), 4,0 (Politikwissenschaft) und 3,8 (Geisteswissenschaften).

[88] Für das Fachteilgebiet Geisteswissenschaften betragen *Recall* und *Precision* beim ersten Testlauf 43, 5 bzw. 26,3 Prozent. Auch hier ist somit ein klarer Anstieg der *Precision* zu verzeichnen. Die intellektuell vergebene Hauptklassifikation wird in 63,2 Prozent der Fälle vergeben. Im Durchschnitt wird sie auf Position 2,7 gerankt.

[89] Die durchschnittliche Anzahl fehlerhafter Klassifikationsnotationen beträgt beim ersten Testlauf 2,0 (Soziologie), 1,3 (Politikwissenschaft) und 1,1 (Geisteswissenschaften).

niedrigsten. Schließlich wird für die Dokumente des Fachteilgebiets Politikwissenschaft nicht nur am häufigsten die Hauptklassifikation gefunden, sondern sie wird durchschnittlich am höchsten gerankt.

In einem nächsten Schritt wurde der sogenannte R-Precision-Wert berechnet. Hierzu wird gleichsam ein *cut-off* Level genau an der Stelle eingeführt, wo auch intellektuell keine weiteren Deskriptoren vergeben wurden. Da eine solche Einstellung technisch nicht möglich ist, wird die Anzahl der automatisch generierten Deskriptoren jeweils im Nachhinein mit der Anzahl der intellektuell vergebenen Deskriptoren gleichgesetzt. Da dies lediglich in denjenigen Fällen möglich ist, wo eine größere oder die gleiche Anzahl an Deskriptoren auch maschinell vergeben wurde, liegt diesem Auswertungsschritt lediglich eine Unterstichprobe zugrunde. *Recall*- und *Precision*-Wert sind in diesen Fällen gleich.

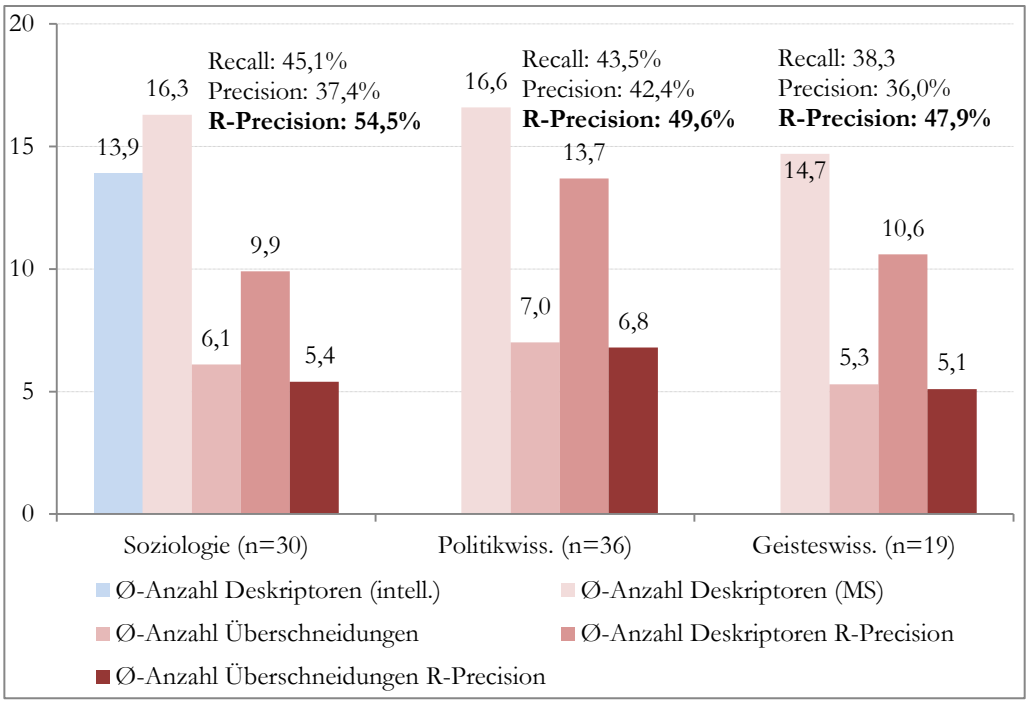


Abbildung 22: Vergleich der R-*Precision*-Werte zwischen den Fachteilgebieten Soziologie, Politik- und Geisteswissenschaften Testlauf III

Im Anschluss wurde für die drei Fachteilgebiete auch eine Auswertung nach den unterschiedlichen Abstract-Arten vorgenommen. Die Diagramme hierzu, in denen für die einzelnen Fachteilgebiete die beiden Testläufe III und IV jeweils gemeinsam ausgewertet werden, finden sich im Anhang (Tab. 13). Nahezu durchweg liegen die Indexierungsergebnisse bei der Vergabe der De-

skriptoren für Dokumente, die anhand eines Autorenreferats erschlossen wurden, näher an der intellektuellen Erschließung als für Dokumente mit Abstracts. Bei der Klassifikation verhält es sich für die Kerngebiete der Datenbank hingegen umgekehrt. Nur für das Fachteilgebiet Geisteswissenschaften erzielt die Indexierungssoftware für Dokumente, die anhand eines Autorenreferats erschlossen wurden, bessere Ergebnisse.

Testlauf IV

Der vierte Testlauf wurde ebenso auf der Grundlage der verschiedenen Trainingsmengen für die Fachteilgebiete Soziologie, Politik- und Geisteswissenschaften durchgeführt. Hierzu wurden zunächst die gleichen Einstellungen wie für den zweiten Testlauf vorgenommen. Verbunden mit den deutlich geringeren Teilmengen im Vergleich zu den ersten beiden Testläufen, führte dies allerdings zu keinen befriedigenden Ergebnissen.[90] Die Einstellungen wurden daraufhin schrittweise bis hin zu einem ausgeglichenen Verhältnis zwischen *Recall* und *Precision* modifiziert. Minimum und Maximum der Trainingsdokumente pro Kategorie wurden auf 20 bzw. 25.000 Dokumente, die minimale Textlänge auf 25 Wörter festgesetzt. Außerdem wurden sowohl für die Klassifikation als auch für die Vergabe der Deskriptoren ein Minimum sowie ein Maximum festgelegt. Für die Klassifikation betrugen diese zwei und fünf Klassifikationen, für die Vergabe der Deskriptoren wurden Minimum und Maximum hingegen sehr hoch angesetzt. Hier wurde sich an der Höchstmarke der intellektuell vergebenen Deskriptoren von 29 orientiert. So betrugen das Minimum 29 und das Maximum 30 Deskriptoren. Mit dieser Einstellung wurde davon ausgegangen, eine größere Unterstichprobe für die Berechnung des R-*Precision*-Wertes zu erzielen.[91]

[90] Die Indexierungssoftware lieferte im Durchschnitt zwischen einem und zwei Deskriptoren. Bei diesen Deskriptoren handelte es sich vor allem um Bezeichnungen, die aufgrund des geographischen Uppostings sehr häufig vergeben werden (z.B. „Entwicklungsland"). Auch konnte die Erhöhung des *Precision*-Wertes beispielsweise dazu führen, dass in einem Indexat zur „Lebenssituation von Jugendlichen und jungen Erwachsenen" in einer vergleichenden Perspektive zwischen Ost- und Westdeutschland lediglich der Deskriptor „neue Bundesländer" vergeben wurde. Der Ähnlichkeitsabgleich mit anderen Dokumenten erfolgte somit in einem zu eng gefassten Bezugsraum.

[91] In den Diagrammen wird erneut versucht, die Unterschiede zwischen drittem und viertem Testlauf zu veranschaulichen.

Ergebnisse des vierten Testlaufs

Wie aus den nachfolgenden Diagrammen (siehe Abb. 23 bis 25) hervorgeht, lässt sich durch ein ausgeglichenes Verhältnis zwischen *Recall* und *Precision* in den Voreinstellungen die Präzision der Indexierungsergebnisse im Vergleich zum vorherigen Testlauf erhöhen. Vor dem Hintergrund der Systemeinstellungen des dritten Testlaufs stellen die Einstellungen des vierten Testlaufs eine Veränderung zugunsten der *Precision* dar.

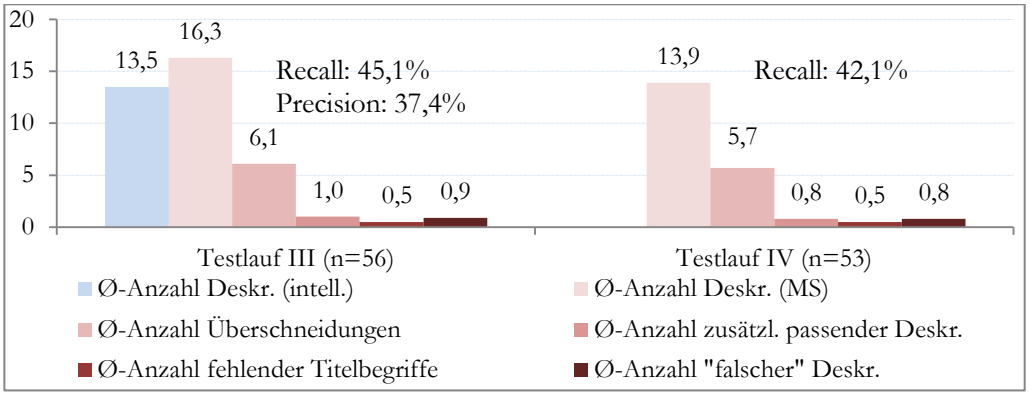


Abbildung 23: Vergleich der Vergabe von Deskriptoren für das Fachteilgebiet Soziologie Testläufe III (n=56) und IV (n=53)

Verbunden mit den Einstellungsveränderungen nehmen sowohl die Anzahl automatisch generierter Deskriptoren als auch die Überschneidungsmenge ab.[92] Hierauf wird das Absinken der *R-Precision* zurückgeführt. Der Ähnlichkeitsabgleich mit anderen Dokumenten aus der Trainingsmenge zur Erkennung des latenten Kontexts vollzieht sich in einem enger gefassten Bezugsraum.

Ähnlich verhält es sich beim Fachteilgebiet Politik. Allerdings steigt hier die Überschneidungsmenge sogar noch geringfügig an. Dies wirkt sich positiv auf den *R-Precision*-Wert aus. *Precision* und *R-Precision* steigen um vier bzw. zweieinhalb Prozentpunkte an (siehe für einen Vergleich der *R-Precision*-Werte zwischen den unterschiedlichen Fachteilgebieten Abb. 29).

[92] Obgleich Minimum und Maximum mit 29 bzw. 30 Deskriptoren sehr hoch angesetzt waren, wurden von der Indexierungssoftware im Durchschnitt deutlich weniger generiert.

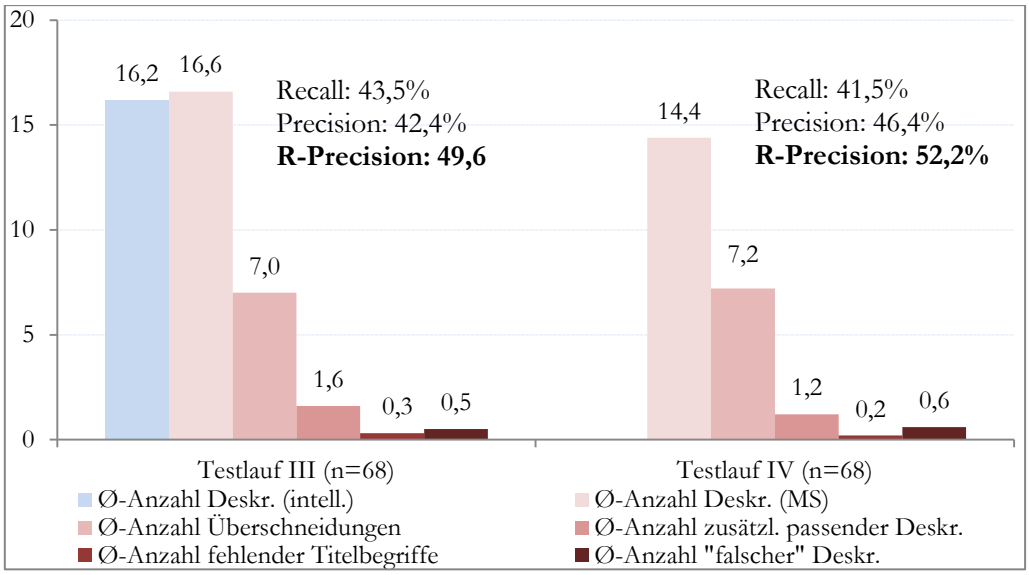


Abbildung 24: Vergleich der Vergabe von Deskriptoren für das Fachteilgebiet Politikwissenschaft Testläufe III (n=68) und IV (n=68)

Noch deutlicher im Verhältnis zu den Werten für Soziologie und Politikwissenschaft ist der Anstieg der *Precision* im Randbereich der Datenbank SOLIS. Zusätzlich nimmt die Anzahl fehlerhafter Deskriptorenvorschläge noch einmal um die Hälfte ab.

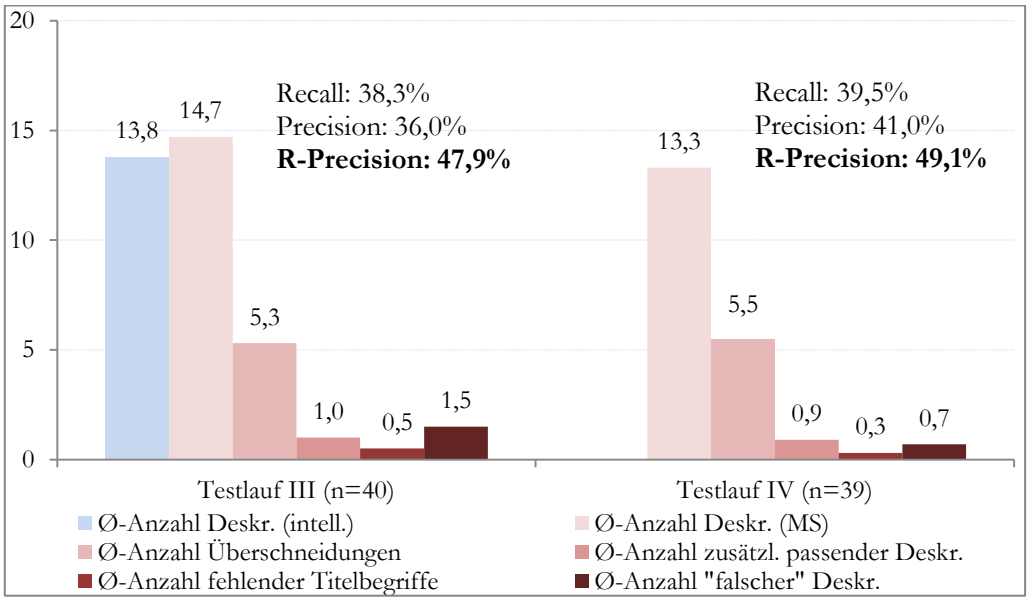


Abbildung 25: Vergleich der Vergabe von Deskriptoren für das Fachteilgebiet Geisteswissenschaften Testläufe III (n=40) und IV (n=39)

Bei der Klassifikation führten die Einführung eines Minimums und eines Maximums bei der Anzahl der Klassifikationszuweisungen zu einer deutlichen Reduktion der Klassifikationsvorschläge. Sowohl in den Kern- als auch im Randbereich wurde dabei die Mindestanzahl an Klassifikationen sogar unterschritten. In den meisten Fällen wurden zwei Klassifikationen vergeben, in

einzelnen Fällen wurden eine oder auch drei Klassifikationen zugewiesen. Die Anzahl der Klassifikationen lag somit sehr nah an der durchschnittlichen Anzahl intellektuell vergebener Klassifikationen (siehe Abb. 26 bis 28).

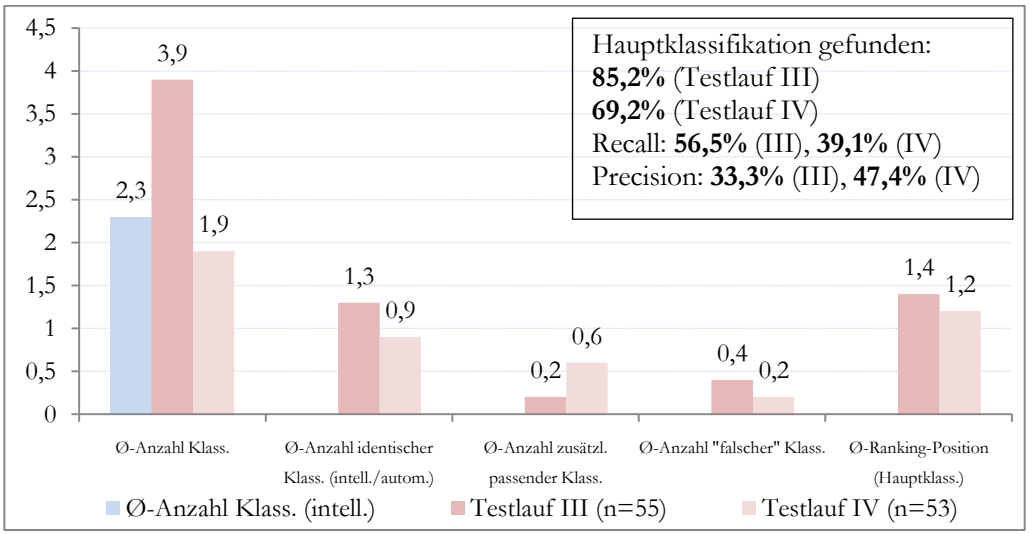


Abbildung 26: Vergleich der Vergabe von Klassifikationen für das Fachteilgebiet Soziologie (Testläufe III (n=55) und IV (n=53)

Für alle drei Fachteilgebiete kann ein Absinken der Überschneidungsmenge festgestellt werden. Dies wirkt sich negativ auf den *Recall* aus. Im Gegenzug kommt es jedoch zu einer klaren Erhöhung der *Precision* für alle drei Fachteilgebiete.

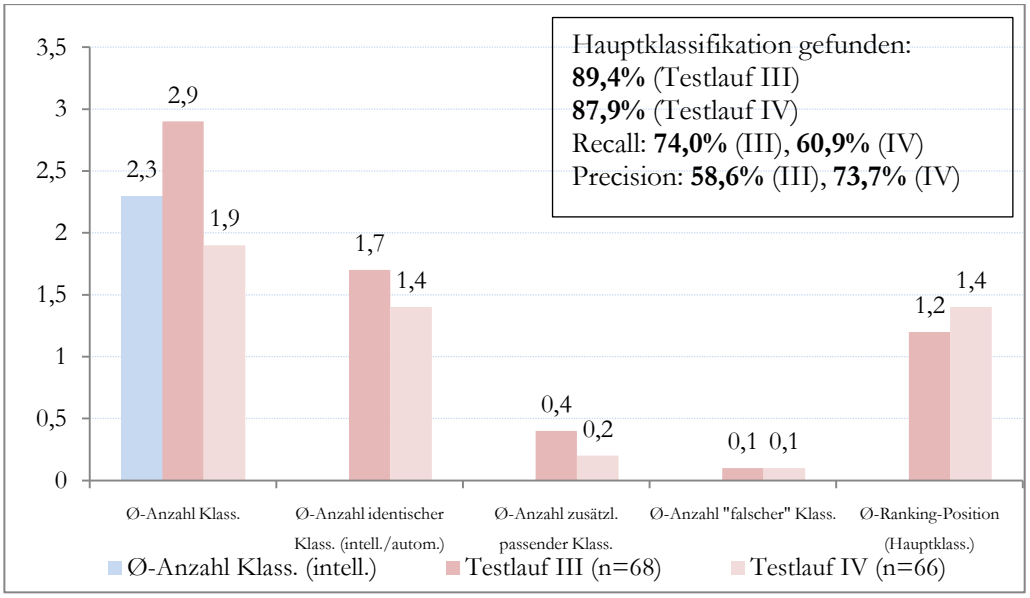


Abbildung 27: Vergleich der Vergabe von Klassifikationen für das Fachteilgebiet Politikwissenschaft Testläufe III (n=68) und IV (n=66)

Auch wird die Hauptklassifikation seltener als beim dritten Testlauf gefunden. Für alle drei Fachteilgebiete geht die Anzahl von Fällen, in denen die

Hauptklassifikation gefunden wird, sogar unter den Wert aus dem ersten Testlauf zurück (vgl. Tab. 3).[93]

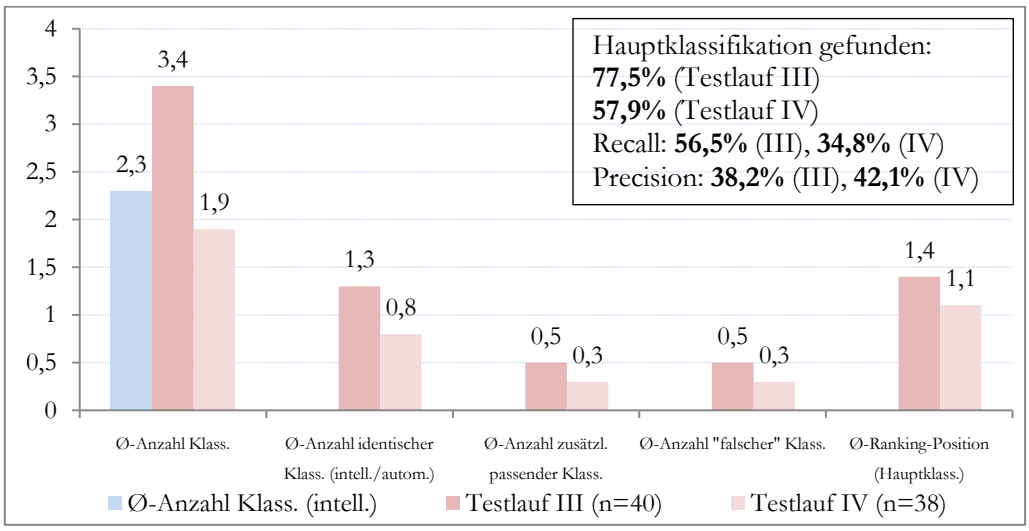


Abbildung 28: Vergleich der Vergabe von Klassifikationen für das Fachteilgebiet Geisteswissenschaften Testläufe III (n=40) und IV (n=38)

Mit den spezifischen Einstellungen des vierten Testlaufs lassen sich erneut die *R-Precision*-Werte berechnen. Allerdings auch erneut nur für eine Unterstichprobe. Die Indexierungssoftware vergibt im Durchschnitt weniger Deskriptoren als die festgesetzten Minimum- und Maximum-Werte.

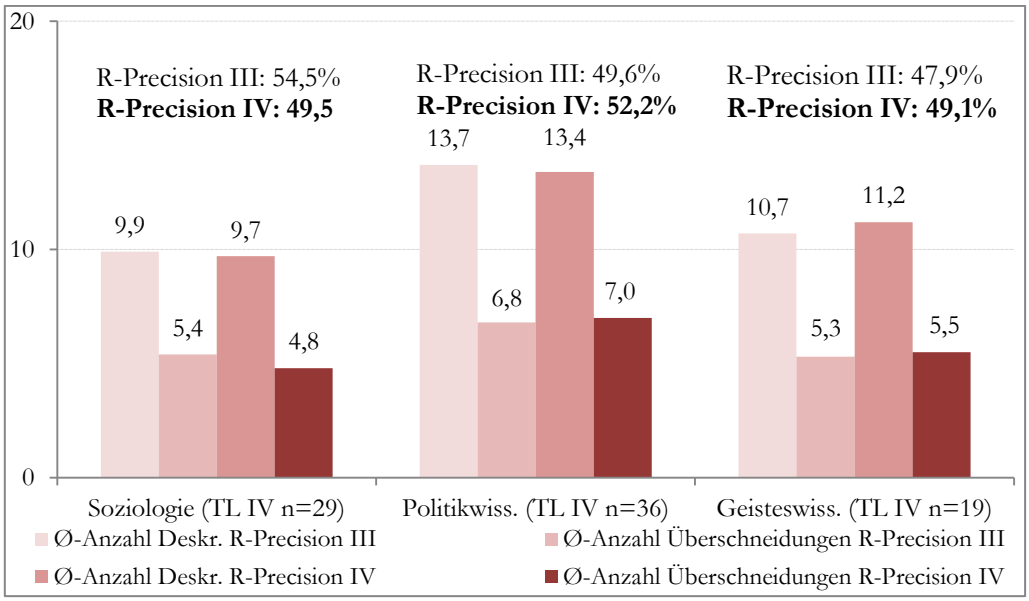


Abbildung 29: Vergleich der R-*Precision*-Werte für die Fachteilgebiete Soziologie, Politik- und Geisteswissenschaften Testläufe III und IV)

[93] Für das Fachteilgebiet Geisteswissenschaften wurde die intellektuell vergebene Hauptklassifikation im ersten Testlauf bei 63,2 Prozent der Dokumente von der Indexierungssoftware generiert.

Mit Blick auf die *R-Precision*-Werte zeichnet sich zunächst sehr deutlich ab, dass sich (theoretisch) mit einer Annäherung der automatisch generierten Anzahl an Deskriptoren an die intellektuell vergebene Deskriptorenanzahl sowohl für den *Recall* als auch für die *Precision* eine deutliche Erhöhung erzielen ließe. Dabei fallen die Unterschiede zwischen drittem und viertem Testlauf für die einzelnen Fachteilgebiete allerdings nicht eindeutig aus. Während sich für das Kerngebiet Politikwissenschaft und den Randbereich Geisteswissenschaften ein Anstieg verzeichnen lässt, ist der *R-Precision*-Wert für das Fachteilgebiet Soziologie deutlich rückläufig. Verbunden mit der im Vergleich zu den beiden anderen Teilgebieten sehr viel größeren Trainingsmenge führt die Erhöhung der *Precision* bei den Systemeinstellungen im Verhältnis zum dritten Testlauf dazu, dass manche intellektuell vergebenen Deskriptoren von der Indexierungssoftware nicht mehr generiert werden.

3.4 Zusammenfassung der Ergebnisse

Ebene der Dokumentenkollektion

Der Aufbau spezifischer MindServer-Versionen für unterschiedliche Fachteilgebiete führt – gemessen an der intellektuellen Erschließung – vor allem zu einer Erhöhung der Präzision der automatischen Indexierung. Mit Übernahme der Standardeinstellungen des MindServer – wie sie bereits beim ersten Testlauf verwendet wurden – wird sowohl für die Kern- als auch für den Randbereich im Vergleich zu den Ergebnissen des ersten Testlaufs ein signifikant höherer *Precision*-Wert erzielt (siehe Tab. 4). Ebenso erhöht sich die Indexierungskonsistenz für alle drei Fachteilgebiete. Durchweg ist sie für das Fachteilgebiet Politkwissenschaft am höchsten. Dies könnte mit einem klarer strukturierten Begriffsapparat bzw. weniger Assoziativrelationen in diesem Bereich des Fachthesaurus Sozialwissenschaften verbunden sein.

	Testlauf I	**Testlauf III**	**Testlauf IV**	***Testlauf II***
Soziologie (n=56)				
Recall	46,6%	45,1%	42,1%	*32,5%*
Precision	31,5%	37,4%	41,1%	*51,4%*
Ind.-Konsistenz	37,6%	40,9%	41,6%	*39,9%*
Politik (n=68)				
Recall	46,0%	43,5%	41,5%	*32,2%*
Precision	36,1%	42,4%	46,4%	*60,1%*

Ind.-Konsistenz	40,0%	42,9%	43,2%	*41,9%*
Geistesw. (n=40)				
Recall	42,0%	38,3%	39,5%	*30,0%*
Precision	32,0%	36,0%	41,0%	*51,4%*
Ind.-Konsistenz	36,3%	37,1%	41,7%	*37,9%*

Tabelle 6: *Recall*- und *Precision*-Werte sowie Indexierungskonsistenz für die Fachteilgebiete Soziologie, Politik- und Geisteswissenschaften Testläufe I bis IV

Verbunden mit der relativen Erhöhung des *Precision*-Wertes bei den Systemeinstellungen für die Vergabe der Deskriptoren beim vierten Testlauf, lässt sich die Präzision der automatisch generierten Indexierungsergebnisse weiter erhöhen. Tendenziell erscheint die Erhöhung in den Kernbereichen der Datenbank etwas deutlicher als im Randbereich.

Diese Erhöhung der Präzision geht mit einer Reduzierung fehlerhafter Deskriptorzuordnungen einher. Jeweils im Vergleich zum ersten Testlauf lässt sich hierdurch zum einen die durchschnittliche Anzahl fehlerhaft vergebener Deskriptoren pro Indexat verringern, zum anderen kann das prozentuale Vorkommen begrifflich falscher Zuordnungen in den fachspezifischen Unterteilmengen reduziert werden.

	Testlauf I	**Testlauf III**	**Testlauf IV**	***Testlauf II***
Soziologie (n=56)				
Durchschnitt	6,2	0,8	0,8	*0,3*
Vorkommen (%)	87,7	44,6	45,3	*20,4*
Politik (n=68)				
Durchschnitt	5,2	0,5	0,6	*0,4*
Vorkommen (%)	84,9	33,0	37,8	*29,3*
Geistesw. (n=40)				
Durchschnitt	3,5	1,5	0,7	*0,7*
Vorkommen (%)	77,5	57,5	43,6	*38,5*

Tabelle 7: Übersicht fehlerhafter Deskriptorzuordnungen für die Fachteilgebiete Soziologie, Politik- und Geisteswissenschaften Testläufe I bis IV

Sowohl bei der Übersicht der *Recall*- und *Precision*-Werte als auch beim Vergleich fehlerhafter Deskriptorzuordnungen stechen die Werte des zweiten Testlaufs hervor. Verbunden mit der stärkeren Gewichtung der *Precision* im Vergleich zum *Recall* um 20 Prozent und der Einführung eines *cut-off* Levels von zehn, erscheinen die Ergebnisse der Indexierungssoftware für sämtliche Werte beim zweiten Testlauf am nächsten an der intellektuellen Erschließung.

Unter alleiniger Berücksichtigung der Werte erscheint die automatische Verschlagwortung mit den Einstellungen des zweiten Testlaufs in Bezug auf sowohl die Präzision der Indexierungsergebnisse als auch die Reduktion fehlerhafter Deskriptorzuordnungen effektiver als über den Aufbau spezifischer Trainingsmengen für die einzelnen Fachteilgebiete. Allerdings geht die starke Erhöhung der Präzision sehr deutlich auf Kosten des *Recall*. Daneben wäre in eine tiefergehenden Untersuchung für den zweiten Testlauf zu klären, inwiefern mit einem *cut-off* Level von zehn vor dem Hintergrund der spezifischen Erschließungsrichtlinien in der sozialwissenschaftlichen Fachinformation auch eine inhaltliche Wiedergabe der dokumentarischen Bezugseinheit gelingt. Denkbar ist etwa, dass in Verbindung mit dem geographischen Upposting vor allem eine geographische Einordnung vorgenommen wird.

Für die *Klassifikation* ergibt sich mit dem Aufbau fachteilgebietsspezifischer MindServer-Versionen für alle drei Fachteilgebiete eine verbesserte Platzierung der intellektuell vergebenen Hauptklassifikation. Am deutlichsten zeigt sich diese im Randbereich der Datenbank. Hier lässt sich bereits mit der Einführung eines *cut-off* Levels, wie sie bereits beim zweiten Testlauf auch für die Klassifikation erfolgte, eine deutlich höhere Ranking-Position der intellektuell vergebenen Hauptklassifikation erzielen.

Daneben lässt sich mit dem Aufbau fachteilgebietsspezifischer Versionen der Indexierungssoftware für den Randbereich auch die Häufigkeit, mit der die intellektuell vergebene Hauptklassifikation im Indexat zugeordnet wird, deutlich erhöhen. Für die Kernbereich trifft dies nur eingeschränkt zu. Gleichwohl bleibt das hohe Niveau, auf dem die intellektuell vergebene Hauptklassifikation von der Indexierungssoftware generiert wird, bestehen. Auch in Bezug auf die Indexierungskonsistenz zeichnet sich für alle drei Fachteilgebiete durch die fachspezifischen Trainingsmengen eine Erhöhung ab. Allerdings erscheint diese für das Fachteilgebiet Soziologie nicht konstant. Aus dem Vergleich der Testläufe III und IV geht ferner hervor, dass sich für alle drei Fachteilgebiete die Einführung eines *cut-off* Levels negativ auf die Klassifikationszuweisung – gemessen an der intellektuellen Klassifikation – auswirkt. Dies zeigt sich bereits bei einem Vergleich der beiden ersten Testläufe.

	Testlauf I	Testlauf II	Testlauf III	Testlauf IV
Soziologie (n=56)				
Hautpklass. gef.	78,2%	78,4%	85,2	69,2
Ranking-Position	1,7	1,6	1,4	1,2
Ind.-Konsitenz	40,5%	45,4%	42,0%	54,3%
Politik (n=68)				
Hauptklass. gef.	92,4%	87,3%	89,4	87,8
Ranking-Position	1,7	1,5	1,2	1,4
Ind.-Konsistenz	58,8%	58,3%	64,4%	66,0%
Geistesw. (n=40)				
Hauptklass. gef.	63,2%	59,0%	77,5	57,9
Ranking-Position	2,7	1,7	1,4	1,1
Ind-Konsistenz	33,9%	37,1%	44,8%	39,5%

Tabelle 8: Vergabe und Ranking-Position der Hauptklassifikation sowie Indexierungskonsistenz für die Fachteilgebiete Soziologie, Politik- und Geisteswissenschaften Testläufe I bis IV

Auch bei der Klassifikation zeigt sich entlang der Testläufe, dass die Anzahl fehlerhafter Klassifikationszuordnungen kontinuierlich deutlich zurückgeht. Durch den Aufbau spezifischer Trainingsmengen für die einzelnen Fachteilgebiete in Kombination mit der Einführung eines *cut-off* Levels (Testlauf IV) geht das Vorkommen fehlerhafter Klassifikationen noch einmal sehr viel deutlicher zurück als allein durch die Einführung eines *cut-off* Levels (Testlauf II).

	Testlauf I	Testlauf II	Testlauf III	Testlauf IV
Soziologie (n=56)				
Durchschnitt	2,0	0,5	0,4	0,2
Vorkommen (%)	76,4	41,2	37,0	21,2
Politik (n=68)				
Durchschnitt	1,3	0,4	0,1	0,1
Vorkommen (%)	62,1	35,0	13,6	11,0
Geistesw. (n=40)				
Durchschnitt	1,1	0,5	0,5	0,3
Vorkommen (%)	60,5	43,6	37,5	23,7

Tabelle 9: Vergabe fehlerhafter Klassifikationen für die Fachteilgebiete Soziologie, Politik- und Geisteswissenschaften Testläufe I bis IV

Dokumentart und Abstract-Art haben keinen signifikanten Einfluss auf das Indexierungsergebnis. Nur geringfügig liegen hier Standardmaße für die Evaluation auseinander. Während bei der Vergabe der Deskriptoren tendenziell die *Recall*- und *Precision*-Werte zu Dokumenten, die anhand eines Autoren-

referats erschlossen wurden, höher liegen, schneidet die Indexierungssoftware bei der Klassifikation derjenigen Dokumente besser ab, die anhand eines Abstracts erschlossen wurden.

Ebene der Einzeldokumente

Für eine abschließende Betrachtung der automatisch generierten Indexierungsergebnisse auf der Dokumentebene gilt es, sich noch einmal bewusst zu machen, dass intellektuelle und maschinelle Indexierung auf unterschiedlichen Ausgangslagen beruhen. Während für den intellektuellen Indexierer das Dokument in der Regel in Autopsie vorliegt, bilden für die Indexierungssoftware einzig Titel und Abstract bzw. Autorenreferat und Inhaltsverzeichnis die Erschließungsgrundlage, über die ein Ähnlichkeitsabgleich zu anderen Dokumenten erfolgt.[94]

Für die Indexierungssoftware können aus der Textgrundlage Schwierigkeiten resultieren, die sich auf die Themen- und Kontexterkennung auswirken. In Bezug auf die Erkennung von Wortbindestrichtilgungen und die Umformung von Adjektiven in Substantive (Derivation) geht aus der Analyse der maschinell erzeugten Indexierungsergebnisse hervor, dass oftmals sowohl Bindestrich- als auch adjektivisch gebrauchte Begriffe von der Software erkannt werden.[95] Gleichwohl ist dies nicht immer der Fall. Hier erscheint es mitunter notwendig, auf der Ebene der einzelnen Dokumente die zutreffenden Bezeichnungen zuzuweisen, damit die Software die korrekte Vergabe der Deskriptoren trainieren kann.[96] Ebenso geht aus der Auswertung der automatisch generierten Erschließungsergebnisse hervor, dass es der Indexierungssoftware vereinzelt nicht gelingt, Begriffe auf ihre Grundform zurückzuführen.[97] Weitaus häufiger hat das automatische Verfahren jedoch Probleme mit der *Kom-*

[94] Die zahlreichen Fußnoten in diesem abschließenden Teil dienen vor allem zur Veranschaulichung.
[95] Als Beispiele lassen sich etwa „armenisch" – generiert den Deskriptor „Armenien" – sowie „wohlfahrtsstaatliche Systeme" – führt zu dem Deskriptor „Wohlfahrt" – anführen. Unter Umständen kann dies auf die latente Kontexterkennung zurückzuführen sein.
[96] In diesem Zusammenhang lässt sich beispielhaft eine Publikation mit dem Begriff „gouvernemental" im Titel anführen. Dieser führt nicht zur Vergabe des Deskriptors „Gouvernementalität". Ebenso wurde der Deskriptor „Neokonservativismus" vermutlich nicht vergeben, da im Autorenreferat die Schreibweise „Neo-Konservativismus" gewählt wurde.
[97] Hierfür exemplarisch ist die Vergabe des Deskriptors „Vertrieb" im Zusammenhang mit Vertriebenen.

positazerlegung.[98] In einzelnen Fällen kann sich dies negativ auf die Vergabe der Geographika auswirken.[99]

Daneben können sich die Wahl bestimmter Formulierungen im Text einerseits und die Auslassung gewisser Begriffe in Titel und Inhaltsbeschreibung andererseits auf das Indexierungsergebnis auswirken. Je nach Kontext können bestimmte Formulierungen mehr oder weniger wahrscheinlich zur Vergabe von Deskriptoren führen. Dies kann mitunter einzig durch ein Verstehen des Abstracts bzw. Autorenreferats oder aber durch die Vermeidung derartiger Formulierungen umgangen werden. In ähnlicher Weise, wie in den Richtlinien zur Abfassung von Abstracts bereits auf den Verzicht bestimmter Formulierungen hingewiesen wird (vgl. GESIS 2005), gilt es, darin weitere Formulierungen aufzunehmen.[100] Im umgekehrten Fall führen Auslassungen dazu, dass bestimmte Deskriptoren, die mitunter für eine regelkonforme Erschließung notwendig sind, trotz Erkennung des latenten Kontexts von der Indexierungssoftware nicht vergeben werden (können).[101]

---

[98] Beispiele hierfür sind „Kriegsfotographie", „Vertrauensverlust" und „Judenhass". Die intellektuell vergebenen Deskriptoren „Fotographie", „Vertrauen" und „Hass" werden von der Indexierungssoftware nicht generiert. Denkbar ist jedoch ebenso, dass die Kompositazerlegung gelingt, die entsprechenden Deskriptoren jedoch nicht vergeben werden, da keine inhaltlich ähnlichen Dokumente in der Trainingsmenge vorhanden sind, die eine Kontexterkennung ermöglichen.

[99] Exemplarisch hierfür sind die Begriffe „Georgienkrieg" und „Nordhessen" im Titel. Sie führen nicht zur Vergabe der Deskriptoren „Georgien" und „Hessen". Im zweiten Fall wird dadurch gegen die Erschließungsrichtlinie zur Vergabe von Geographika verstoßen, wonach – wenn möglich – für Deutschland eine Präzisierung des geographischen Bezugsraums bis auf Ebene der Bundesländer vorgenommen werden soll.

[100] Neben der Vermeidung verneinender Formulierungen („unabhängig von Raum und Zeit", oder: „nicht nur Gegenstand von Kulturmanagement und Kulturpolitik"), wie sie als Hinweis zum Abfassen eines Abstracts bereits Eingang in die Indexierungsrichtlinien gefunden haben, zählen dazu etwa Formulierungen wie „jenseits von" oder Begriffe und Bezeichnungen in Anführungszeichen oder Klammern („die Rezeption der als 'westlicher Neomarxismus' verstandenen kritischen Theorie"). Derartige diffizile Bedeutungsfacetten werden von der Indexierungssoftware nicht erkannt. Hier erscheint allerdings der Hinweis angebracht, dass ein solches Vorgehen ausschließlich beim Anfertigen von Abstracts zum Einsatz kommen kann. Bei der Verwendung von Autorenreferaten und Inhaltsverzeichnissen als Erschließungsgrundlage, wie sie in der sozialwissenschaftlichen Fachinformation bereits ausschließlich etwa bei der Indexierung von Gesamtaufnahmen von Sammelwerken genutzt werden, ist die Einflussnahme auf die Wahl der Formulierung nur sehr eingeschränkt bzw. gar nicht möglich.

[101] Ein Beispiel bildet der Deskriptor „Bundesrepublik Deutschland". In Publikationen, die einen Bezug zu Deutschland aufweisen, wird dieses Geographikum intellektuell vergeben. Finden sich allerdings keine Formulierungen in Titel und Inhaltstext, die auf diese geographische Einordnung hinweisen, und legt auch der Ähnlichkeitsabgleich mit anderen Dokumenten aus der Trainingsmenge eine Vergabe dieses Deskriptors nicht nahe, wird er in der Regel auch nicht vergeben. Dass gerade dieser Deskriptor nicht vergeben wird, hängt mit der inver-

Die Kontexterkennung einer Publikation wirkt sich deutlich positiv auf die automatisch generierte Indexierung aus. Sie begünstigt, dass implizite Inhalte einer Publikation berücksichtigt werden, die nicht lexikalisch in Titel und Inhaltstext aufgeführt sind. Vor allem für die Erschließung auf der Grundlage von Autorenreferaten, die sich nicht so sehr am Thesaurusvokabular orientieren, ist dies für das maschinelle Indexierungsergebnis – gemessen am intellektuell generierten Indexat – von Vorteil.[102]

Ein weiterer Vorteil im Zusammenhang mit der Erkennung des latenten Kontexts einer Publikation besteht darin, dass sehr spezielle Begriffe, so etwa Personennamen, obgleich sie im Titel oder Inhaltstext aufgeführt werden, bei der Vergabe der Deskriptoren gerade *keine* Berücksichtigung finden. Trotz eines hohen Diskriminanzeffekts werden sie nicht als Schlagwort generiert.[103] Nachteilig wirkt es sich im Zusammenhang mit der Kontexterkennung jedoch aus, wenn Publikationen ein Thema aus einer sehr speziellen Perspektive behandeln, oder eine sehr spezielle Eingrenzung des Themas vornehmen. Hier ist verfahrenstechnisch bedingt – sowie je nach Trainingskollektion noch einmal unterschiedlich – eine erfolgreiche Kontexterkennung mitunter nur sehr eingeschränkt möglich.[104] Ähnlich verhält es sich bei Publikationen, die sich durch ungleich gewichtete Themenaspekte oder eine Vielzahl unterschiedlicher Themenfelder auszeichnen. Dominiert in diesen Fällen etwa ein

---

sen Dokumenthäufigkeit des Terms zusammen. So bildet er aufgrund der Regeldefinition für die intellektuelle Erschließung einen der am häufigsten in der Datenkollektion vergebenen Deskriptor. Ein weiteres Beispiel bildet eine Publikation zu Zentralamerika. Während im Abstract keinerlei Ländernamen aufgeführt sind, wurden intellektuell die Geographika „Belize“, „Costa Rica“, „Honduras“, „Nicaragua“ und „Panama“ vergeben.

[102] Exemplarisch hierfür ist die Vergabe des Deskriptors „Weltgeschichte“, obgleich im Autorenreferat stets der Begriff „Globalgeschichte“ verwendet wird.

[103] Ein Beispiel bildet eine Publikation mit dem Titel „Thomas Manns Albtraum. Potential und Paradoxien europäischer Erinnerungspolitik“, in der es um das Thema der Vergangenheitsbewältigung geht. „Thomas Mann“ wird nicht als Deskriptor generiert. Generell ist dies bei Titelformulierungen, die eher Aufmerksamkeit erzeugen sollen, als bereits über den Inhalt der Publikation Auskunft zu geben, von Vorteil.

[104] Exemplarisch hierfür ist die Behandlung eines prominenten Themas unter einer besonderen Länderperspektive. So wurde zu einer Publikation mit dem Titel „Die Beziehung zwischen Vertrauen und Führungsstil in slovenischen[sic] Organisationen“ in keinem der vier Testläufe der Deskriptor Slowenien vergeben. Eingehender betrachtet kann dies mehrere Gründe haben. Zum einen handelt es sich hier um einen orthographischen Fehler. Zum anderen kann es sein, dass in diesem Fall die Umformung des Adjektivs in das Substantiv Slowakei nicht funktioniert. Schließlich kann es aber auch damit zusammenhängen, dass andere Dokumente, die in der Trainingsmenge manuell mit dem Deskriptor „Slowakei“ versehen wurden, nicht von Vertrauen und Führungsstil handeln, was eine adäquate Kontexterkennung nahezu unmöglich macht.

Themenbereich sehr deutlich oder ist von einem anderen ebenfalls behandelten Kontextbezug sehr weit entfernt – dies kann sich etwa darin äußern, dass die unterschiedlichen Aspekte nur selten auch in anderen Dokumenteinheiten gemeinsam thematisiert werden – kann es sein, dass die Inhaltsaspekte, die weniger im Vordergrund stehen, bei der Schlagwortvergabe keine Berücksichtigung finden. Die Vielzahl an Aspekten, die mitunter nicht erkannt wird, kann gerade bei interdisziplinären Forschungsarbeiten ein Problem darstellen, so etwa bevorzugt bei Gesamtaufnahmen von Sammelwerken oder bei Essay-Sammlungen (eines Verfassers), die Aufsätze zu sehr unterschiedlichen Themenbereichen beinhalten.[105]

Werden auf der Grundlage von Titel und Abstract im Zuge der Kontexterkennung jedoch sehr spezifische Kontextrelationen aufgebaut, so werden von der Indexierungssoftware je nach Einzelfall auch andere Begriffe aus dem semantischen Umfeld erzeugt und als Deskriptoren hinterlegt.[106] Durch eine Regeldefinition ließe sich in diesen Fällen vermeiden, dass automatisch bestimmte Bezeichnungen aus dem begrifflichen Umfeld generiert werden. In eine ähnliche Richtung weist, wenn zu einem bestimmten Themengebiet sehr spezifische Begriffe im Inhaltstext aufgeführt werden, die dann auch von der Indexierungssoftware aufgrund ihres hohen Diskriminanzeffekts als Deskriptoren vergeben werden.[107] Handelt es sich hierbei um Ländernamen, kann dies zu zusätzlichem Ballast führen. So wird regelkonform mitunter zusätzlich ein geographisches Upposting vorgenommen.

Mitunter verläuft die Erkennung des weitergefassten Kontexts allerdings nicht erfolgreich. Bestimmte Terme im Text können etwa Bestandteile von

[105] Ein Beispiel bildet eine Essay-Sammlung, deren Autorenreferat Begriffe wie „Mobilisierungs- und Integrationsideologie“ ebenso beinhaltet wie „Nationalismus“, „Puritanismus“, „Preußennostalgie“, „Marktgesellschaft“ und „Wirtschaftskrise“. In fast allen Testläufen führt dies einzig zu den Deskriptoren „Bundesrepublik Deutschland“ und „Nationalismus“. Dieses Problem kann umso stärker auftreten, je deutlicher das Verhältnis zwischen *Precision* und *Recall* zugunsten der *Precision* eingestellt ist.

[106] Ein Beispiel bildet eine Publikation zur politischen Theorie und Ideengeschichte. Die Erwähnung von Alexis de Tocquevilles im Abstract führt dazu, dass die Indexierungssoftware zusätzlich die Deskriptoren „Hobbes, T.“ sowie „Machiavellismus“ generiert.

[107] Hier lässt sich beispielhaft die Vergabe mehrerer Fernsehanstalten („ARD“, „ZDF“ und „RTL“) anführen, die im Abstract in Klammern angegeben sind. Intellektuell wurden diese nicht vergeben.

Bezeichnungen bilden, und fälschlicherweise mit ihnen assoziiert werden.[108] In anderen Fällen werden von der Indexierungssoftware Begriffe erkannt, die im Thesaurus allerdings nur in Kombination mit anderen Begriffen bzw. Begriffsbestandteilen vergeben werden können. Der zusammengesetzte Begriff entspricht mitunter nicht mehr dem entsprechenden semantischen Umfeld des Inhaltstextes.[109]

Eher negativ auf die Bewertung der Indexierungsergebnisse wirkt es sich aus, wenn durch die Kontexterkennung eine ganze Reihe von Bezeichnungen des entsprechenden semantischen Feldes generiert wird. Dieses Verfahren kann einerseits die Wahrscheinlichkeit erhöhen, dass automatisch eine bestimmte Bezeichnung vergeben wird, die auch intellektuell gewählt würde. In diesem Sinne kann es das Evaluationsergebnis – wird es einzig gemessen an der Kongruenz mit der intellektuellen Erschließung – verbessern, indem es zu einer Erhöhung des *Recall* führt. Nicht unbedingt muss jedoch der entsprechende intellektuell vergebene Deskriptor von der Indexierungssoftware hierbei auch generiert werden. Andererseits kann es sich negativ auf die *Precision* der Indexierungsergebnisse auswirken, da unter Umständen eine größere Anzahl an Deskriptoren generiert wird, die nicht intellektuell vergeben wurden.[110] Vor dem Hintergrund des konkreten Anwendungsfalls, bei dem es um die Implementierung einer prozessunterstützenden automatischen Erschließung geht, würde dies zu weiteren Deskriptorvorschlägen führen, unter denen der Indexierer eine Auswahl zu treffen hätte, um die Gesamtmenge der Deskriptoren einzugrenzen.[111] Konträr dazu werden in einzelnen Fällen gar keine Deskriptoren generiert, obgleich die formalen Voraussetzungen entspre-

[108] Ein Beispiel hierfür bilden die Begriffe „Bürger“ und „Bürgergesellschaft“, die zur Vergabe des Deskriptors „bürgerliche Gesellschaft“ führen.
[109] So führen die Titelbegriffe „Wohnmobilität“, „Mobilitätsraten“ und „Mobilitätsgründe“ etwa zu dem Deskriptor „soziale Mobilität“. „Mobilität“ wird nach dem Thesaurus Sozialwissenschaften nicht allein vergeben. Stattdessen finden sich darunter die Vorzugsbenennungen „Berufsmobilität“, „horizontale Mobilität“, „Migration“, „regionale Mobilität“, „soziale Mobilität“ und „vertikale Mobilität“.
[110] Ein Beispiel hierfür ist die Vergabe der Deskriptoren „Bürgerbeteiligung“, „direkte Demokratie“, „Volksabstimmung“, „Volksbegehren“ und „Volksentscheid“ zu einer Publikation über direktdemokratische Entscheidungsverfahren. Intellektuell wurden hier einzig „direkte Demokratie“ und „Bürgerbeteiligung“ vergeben.
[111] Bei einem direkten Einspielen der automatisch generierten Deskriptoren in die Datenbank, hätte dies deutliche Auswirkungen auf das Information Retrieval. Für die Recherche ergäben sich hierdurch zusätzliche Sucheinstiege. Diese ließen sich im konkreten Anwendungsfall allerdings durch den Vorschlagsdienst weiterer Schlagwörter aus dem kontrollierten Vokabular erneut sinnvoll eingrenzen.

chend der gewählten Systemeinstellungen erfüllt sind. Dies wurde darauf zurückgeführt, dass auf der Grundlage der Trainingsmenge keine, respektive keine eindeutige Kontext- bzw. Ähnlichkeitserkennung zu anderen Dokumenten stattfinden konnte.

Der höhere *Recall* in den Testläufen I und III führt dazu, dass die Kontexterkennung in einem weiter gefassten Bezugsraum erfolgt. Dadurch können stärker latente bzw. weniger eng gefasste Beziehungen zu anderen Dokumenten ermittelt werden. Die weiter gefasste Kontexterkennung äußert sich etwa darin, dass Komposita, die in Titel oder Inhaltstext aufgeführt sind, zerlegt werden. So werden zum Teil allgemeinere Begriffe gewählt als im anschließenden vierten Testlauf.[112] Sie kann sich aber auch in Unterschieden in der zeitlichen Einordnung der Dokumente zeigen.[113]

Latente Begriffsbeziehungen bzw. Kontextbezüge werden häufiger aufgefunden. Dies kann dazu führen, dass mitunter ein engeres Schlagwort generiert wird, als es alleine aus der präzisen Berücksichtigung der Begriffe aus Titel und Abstract möglich ist.[114] Eine weiter gefasste Kontexterkennung kann somit unter Umständen zur Vergabe eines engeren Schlagwortes führen.[115]

---

[112] Ein Beispiel bildet die Zerlegung des Kompositums „Entscheidungsprozess“, die im Autorenreferat zweimal im Plural verwendet wird, in „Entscheidung“ und „Prozess“. Durch die Veränderung der Systemeinstellungen im vierten Testlauf zugunsten der *Precision*, wird dies – wie bereits vom ersten zum zweiten Testlauf – wieder rückgängig gemacht. Dadurch wird eine Übereinstimmung mit der intellektuellen Indexierung erzielt. Hier wurde ebenso der Deskriptor „Entscheidungsprozess“ vergeben.

[113] Diese Auswirkungen der Einstellungsveränderungen auf die Testergebnisse zeigen sich anschaulich bei den Indexierungsergebnissen zu einer Publikation über Konsumgewohnheiten des aufkommenden Bürgertums im 18. Jahrhundert. Während durch die hohen *Recall*-Werte bei den Testläufen I und III neben den Zeitschlagwörtern „18.-“, „19.-“ und „20. Jahrhundert“ ebenso „17. Jahrhundert“ vergeben wird, generiert der vierte Testlauf lediglich „18.-“ bis „20. Jahrhundert“. Der zweite Testlauf hingegen – verbunden mit der Einführung eines *cut-off* Levels von 10 und dem relativ hohen *Precision*-Wert – generiert lediglich „19. Jahrhundert“, obgleich keine explizite Erwähnung des Zeitbegriffs „19. Jahrhundert“ im Text erfolgt. Der enger gezogene Bezugsraum zur Ermittlung des Kontextes zeigt sich auch bei einer Publikation zu „Dienstmädchen um 1900“. Während in Testlauf IV einzig „18.-“ und „19. Jahrhundert“ vergeben werden, finden sich im Indexierungsergebnis des dritten Testlaufs zusätzlich ebenso „17.-“ und „20. Jahrhundert“. Die fachteilgebietsspezifische Einordnung sorgt hierbei dafür, dass – anders als im ersten Testlauf – auch „17. Jahrhundert“ generiert wurde.

[114] Ein Beispiel bildet die Vergabe des Deskriptors „historische Analyse“ im dritten Testlauf, obgleich im Inhaltstext lediglich der Begriff „Analyse“ verwendet wird.

[115] Dass dies nicht immer der Fall ist, sondern jeweils von der Textgrundlage abhängig ist, belegt etwa ein Abstract, in dem dreimal der Begriff „Öffentlichkeitsarbeit“ verwendet wird. Während dies beim vierten Testlauf erneut zur Vergabe des gleichnamigen Deskriptors führt, werden beim dritten Testlauf für diesen Aspekt die Schlagwörter „Kommunikation“ und „öffentliche Kommunikation“ vergeben.

Umgekehrt kann eine stärkere Berücksichtigung von Titel und Inhaltstext, wie sie durch einen höheren *Precision*-Wert erzielt wird, zu einem weiter gefassten Schlagwort führen.[116] Entscheidend ist jeweils die Textgrundlage und wie sie gewichtet wird. Daneben kann eine weniger strenge Kontexterkennung allerdings auch zu fehlerhaften Begriffsbeziehungen führen.[117]

Durch eine Erhöhung der *Precision*, wie sie vor allem in Testlauf II aber relativ zu den Testläufen I und III auch im vierten Testlauf erfolgte, findet der Ähnlichkeitsabgleich zu anderen Dokumenten hingegen in einem enger gefassten Kontextraum statt. Weiter entfernt liegende Kontextbezüge werden nicht (mehr) erkannt.[118] Dies kann zur Vergabe sehr spezifischer Begriffe führen und sich in dieser Weise etwa positiv auf die Erkennung von Geographika auswirken,[119] aber auch zu einem fehlerhaften Ähnlichkeitsbezug führen.[120] Damit verbunden ist auch eine andere Gewichtung der Termfrequenz. Seltener aufgeführte Begriffe erzielen einen höheren Konfidenzwert, häufiger genannte Begriffe werden unter Umständen nicht mehr als Deskriptor generiert.[121]

---

[116] Ein engerer Kontextraum führt beispielweise eher zu „Modularisierung". Dieser Begriff wird im Abstract auch zweimal verwendet. Ein weiter gefasster Kontextraum führt hingegen zu dem Deskriptor „Bologna-Prozess". Im ersten Testlauf wird beides vergeben. Im zweiten nur „Modularisierung".

[117] Exemplarisch lässt sich die Vergabe des Deskriptors „Dialekt" bei einer Publikation mit dem Titel „Dialektik der atomaren Abschreckung" anführen. Da diese fehlerhafte Zuordnung beim ersten Testlauf nicht erfolgt, hängt sie auch mit dem Aufbau der fachteilgebietsspezifischen MindServer-Versionen zusammen. In diesem Zusammenhang ist auch eine fehlerhafte Grundformzerlegung, die durch eine Regeldefinition behoben werden könnte, sehr wahrscheinlich.

[118] Exemplarisch lässt sich das Indexat zu einer Publikation mit dem Titel „Akteure der Vernichtung. Deutsche und sowjetische Täter – ein Vergleich" anführen. Das Geographikum „UDSSR", das einzig durch das Adjektiv „sowjetisch" sowie durch die Verwendung des Begriffs „Stalinismus" (in)direkt im Abstract vorkommt, wird einzig im dritten Testlauf generiert. Hieran macht sich beispielhaft somit ebenso die fachteilgebietsspezifische Eingrenzung positiv bemerkbar.

[119] Ein Beispiel bildet hierfür etwa der Titelbegriff „Nepal". Dieses Geographikum wurde lediglich im zweiten und im vierten Testlauf vergeben. Verbunden mit der erhöhten *Precision* im zweiten Testlauf betrug der Konfidenzwert hier 0.99, beim vierten Testlauf hingegen 0.63.

[120] So wird etwa ein Dokument, in dem die Begriffe Bundesstaat und Staatenbund sowohl im Titel als auch im Abstract aufgeführt werden, mit dem Deskriptor „USA" erschlossen, obgleich die Publikation keinerlei Bezug dazu aufweist. Unter zwölf vergebenen Schlagwörtern erhält dieser Deskriptor den höchsten Konfidenzwert.

[121] Auch hierfür lassen sich mehrere Beispiele anführen. So werden in einem Abstract etwa siebenmal der Begriff „Volksbegehren" und dreimal der Begriff „Volksentscheid" verwendet. Während dies beim dritten Testlauf zu einem Konfidenzwert von 0.99 für den Deskriptor „Volksbegehren" und zu einem Konfidenzwert von 0.87 für das Schlagwort „Volksentscheid" führt, erzielt letzterer beim vierten Testlauf noch einen Konfidenzwert von 0.78.

Die fachteilgebietsspezifischen Versionen der Indexierungssoftware generieren vor dem Hintergrund der fachlich spezifizierten Trainingsmengen zum Teil enger gefasste Zuordnungen.[122] Dies entspricht allerdings nicht unbedingt, so soll noch einmal betont werden, auch der intellektuellen Indexierung. Einzelfallspezifisch kann ein enger gefasster Bezugsraum zur Kontexterkennung auch zu einer sehr stark verkürzten Inhaltswiedergabe oder zu fehlerhaften Ähnlichkeitsverweisungen führen.[123] Zum Teil wird durch die Eingrenzung der Trainingsmenge auf ein Fachteilgebiet aber auch erst eine Kontexterkennung möglich, die stärker einer intellektuellen Indexierung entspricht. So werden manche intellektuell vergebenen Deskriptoren erst in den beiden letzten Testläufen generiert.[124]

Des Weiteren wurden intellektuell teilweise Deskriptoren vergeben, zu denen bislang nur sehr wenige Dokumente in der Trainingsmenge vorlagen. In einem solchen Fall ist es der Indexierungssoftware je nach Systemeinstellungen nur bedingt möglich, die entsprechenden Deskriptoren zu generieren. In anderen Fällen wurden intellektuell Deskriptoren vergeben, die in der Version des Thesaurus, die dem System für die Vergabe von Deskriptoren für die Testdokumente zur Verfügung gestellt wurde, noch nicht enthalten waren.

Einen weiteren Problembereich bildet die mangelnde Trennschärfe bestimmter Deskriptoren. Obgleich im Thesaurus die Verwendung der einzelnen Be-

---

„Volksbegehren“ wird hingegen von der Indexierungssoftware gar nicht mehr als Deskriptor generiert. Hierzu lässt sich auch ein Indexat zu einem Dokument mit dem Titel „Friedhofsgespräche. Der Friedhof als Ort der Kommunikation“ anführen. Obgleich, bzw. vermutlich gerade weil, „Friedhof“ noch ein weiteres Mal im Abstract aufgeführt ist, wird dieser Begriff beim vierten Testlauf nicht mehr als Deskriptor generiert, während er beim dritten Testlauf noch einen Konfidenzwert von 0.99 erzielte.

[122] Ein Beispiel ist die Vergabe des Deskriptors „internationale Wirtschaftsbeziehungen“ im dritten Testlauf anstelle von „internationale Beziehungen“ im ersten Testlauf. Ein weiteres Beispiel ist die Verknüpfung bzw. Assoziation der Begriffe „Untergebene“, „Führungstypen“ und „Führungsstil“ sowie „Kontrolle“ mit dem Begriff „Oligarchie“ in einem Autorenreferat zu einer Publikation im Bereich Organisationssoziologie und Personalwesen im vierten Testlauf. In einem anderen Fall findet eine Verknüpfung des Themas „Historikerstreit“ mit „Nationalsozialismus“ einzig in Testlauf IV statt. Hieran zeigt sich ein positiver Effekt der fachteilgebietsspezifischen Eingrenzung der Trainingsmenge.

[123] Ein Beispiel ist die alleinige Vergabe der inhaltlichen Deskriptoren „Islamismus“ und „Terrorismus“ zu einer Publikation mit dem Titel „Islam und Islamismus in der Türkei“ im vierten Testlauf. Die veränderten Systemeinstellungen zwischen den beiden letzten Testläufen führen dazu, dass sich die Anzahl der Deskriptoren um 20 Schlagwörter reduziert.

[124] Exemplarisch ist der Deskriptor „Frau“ in einer Publikation zur Rettung von Juden in der Zeit des Nationalsozialismus durch eine Bordellbetreiberin in dessen Abstract der Begriff „Frau“ gar nicht erwähnt wird. Zu einer anderen Publikation wurde der Titelbegriff „Lernort“ erst beim vierten Testlauf generiert.

griffe expliziert wird, weisen die Dokumente, die intellektuell mit bestimmten Schlagwörtern erschlossen wurden, inhaltlich mitunter deutliche Überschneidungen zu ähnlichen semantischen Feldern auf. Für die Indexierungssoftware kann dies die Auswahl der gemäß dem Thesaurus genau passenden Deskriptoren erschweren.[125] Es gelingt der Indexierungssoftware auf der Grundlage der Trainingsmenge nicht, bestimmte Begriffsbedeutungen ebenso eng zu ziehen, wie dies im kontrollierten Vokabular vorgesehen und hinterlegt ist.

In Bezug auf die Vergabe der Klassifikationsnotationen zeigt sich, dass die Indexierungssoftware oftmals sehr viele Klassifikationsvorschläge generiert. Vor allem bei den Standardeinstellungen der Indexierungssoftware kommt es zu einer Vielzahl an Klassifikationsvorschlägen, die zum Großteil mitunter zu einer Klassifikationsgruppe gehören. Dies hängt damit zusammen, dass sich eine Fülle von Publikationen sehr unterschiedlichen Wissensgebieten oder mehreren Unterdisziplinen einer Domäne zuordnen lässt. Mitunter führt die Vielzahl an Aspekten in einer Publikation schließlich dazu, dass keinerlei *K*lassifikation vergeben wird, obgleich die formalen Voraussetzungen dafür erfüllt werden. Gleiches lässt sich für Publikationen beobachten, die sehr spezielle Themen behandeln.[126]

Um eine hohe Anzahl an Klassifikationen zu vermeiden, ist es daher sinnvoll, für die Klassifikation einen *cut-off* Level einzuführen. Eine Evaluation anhand der Ranking-Position der Hauptklassifikation gibt allerdings die Indexierungsqualität unter Umständen nicht vollständig wieder. So kann es sich bei einer hoch gerankten maschinell generierten Klassifikation etwa um eine der intellektuell generierten Nebenklassifikationen handeln. Es muss sich somit auch nicht um ein unbefriedigendes Indexierungsergebnis handeln, wenn die intellektuell vergebene Hauptklassifikation nicht generiert wurde. Hierbei ist ebenso zu berücksichtigen, dass bei einem niedrigen *cut-off* Level von der Indexierungssoftware mitunter nur eine Klassifikation vergeben wird.

[125] Beispiele hierfür sind etwa die Deskriptoren „ökonomische Entwicklung“ und „Wirtschaftsentwicklung“, „Konfliktregelung“ und „Konfliktbewältigung“ sowie „Berufsschule“ und „berufsbildende Schule“. Im Thesaurus Sozialwissenschaften finden sich aufgrund des weit gefassten Gegenstandsbereichs zahlreiche Begriffspaare, die inhaltlich deutliche Überschneidungen aufweisen.

[126] Als Beispiel lässt sich ein Aufsatz zur „Freimaurerei in Briefen Johann Gottlieb Fichtes und Theodors von Schön“ nennen. In keinem der vier Testläufe wird hierzu eine Klassifikation ermittelt.

Eine Ermittlung der Überschneidungswerte ist auch aus diesem Grund sinnvoll.[127]

[127] Ein Beispiel ist die intellektuelle Vergabe der Klassifikationen „politische Willensbildung“ als Haupt- und „Friedens- und Konfliktforschung“ als Nebenklassifikation. In den Testläufen III und IV wird ausschließlich die Nebenklassifikation vergeben.

Teil IV: Diskussion und Ausblick

Die vorliegende Studie zeichnet sich durch ihren engen Anwendungsbezug und die Evaluation der Indexierungssoftware bis auf die Dokumentenebene aus. Hierin besteht das besondere Anliegen der Arbeit. Abschließend werden zentrale Ergebnisse der Evaluationsstudie zusammengefasst und weiterführende Fragestellungen angedeutet. Schließlich werden bestimmte Problematiken in Bezug auf die Evaluation der Indexierungsergebnisse reflektiert.

Aus der Evaluationsstudie geht als eine zentrale Erkenntnis hervor, dass die automatisch generierten Indexierungsergebnisse für die Kerngebiete der Datenbank SOLIS tendenziell eine höhere Übereinstimmung mit dem intellektuellen Indexat aufweisen als die Ergebnisse für die Randgebiete. Eine automatische Vor-Indexierung für die Randgebiete würde im Vergleich zu den Kerngebieten vor allem weniger vollständige Indexierungsvorschläge liefern. Dies erklärt sich aus der niedrigeren Gesamtzahl an Dokumenten zu den Randbereichen in der Dokumentenkollektion, wodurch die Kontext- bzw. Ähnlichkeitserkennung zu anderen Dokumenten erschwert wird. Verbunden mit der durchschnittlich niedrigeren Gesamtzahl an Deskriptoren pro Dokument wurden je nach Systemeinstellungen mitunter allerdings für die Randbereiche ebenso weniger inhaltlich fehlerhafte Deskriptoren vergeben als für die Kerngebiete. Insgesamt zeigen sich zwischen den bisher betrachteten Randgebieten jedoch deutliche Unterschiede. In weiteren Testläufen ließen sich die Unterschiede beim Erschließungsverhalten der Indexierungssoftware zwischen den einzelnen Randbereichen genauer untersuchen. Hierbei ließe sich darauf zurückgreifen, dass bereits aktuell eine Vor-Indexierung durch die Software MindServer technisch möglich ist und die Ergebnisse gespeichert werden können. Diese Indexierungsergebnisse ließen sich nach den einzelnen intellektuell vergebenen Hauptklassifikationen auswerten.

Aus der Evaluationsstudie geht ferner hervor, dass sich durch die Einführung eines *cut-off* Levels die Präzision der automatisch generierten Indexierungsvorschläge – gemessen an der intellektuellen Erschließung – deutlich erhöhen und die Anzahl fehlerhafter Kategorisierungsvorschläge sehr stark reduzieren lässt. Dies gilt vor allem für die Vergabe der Deskriptoren. Auf die Indexie-

rungskonsistenz hat dies nur geringen Einfluss. Sie liegt bezogen auf die Gesamtstichprobe beim zweiten Testlauf nur unwesentlich höher als bei den anderen Testläufen. Beim Einsatz einer semiautomatischen Indexierung sollte somit – auch um für den Indexierer den zeitlichen Aufwand bei der Auswertung des automatisch generierten Indexats gering zu halten – die Anzahl sowohl der automatisch vergebenen Deskriptoren als auch der Klassifikationsnotationen eingeschränkt werden.[128]

Auch hier zeigen sich jedoch deutliche Unterschiede zwischen den einzelnen Fachteilgebieten, die entsprechend der Klassifikation Sozialwissenschaften in der Datenbank SOLIS enthalten sind. Während sich mit der Einführung eines *cut-off* Levels für die Kernbereiche von einer Erhöhung der Präzision ausgehen lässt, kann sie für einzelne Randbereiche sogar einen Rückgang bedeuten. Aufgrund der geringen Dokumentenzahl sind hierzu erneut noch keine allgemeingültigen Aussagen möglich.

Ein Aufbau fachteilgebietsspezifischer Kontext- und Konzepträume wirkt sich ebenso positiv auf die Präzision der Indexierungsergebnisse – gemessen an den intellektuell generierten Vergleichsdaten – aus. Die dadurch erzielte größere Homogenität der Trainingsmengen erleichtert die Ähnlichkeitserkennung zu anderen Dokumenten. Die Erhöhung der Präzision bleibt jedoch hinter dem *Precision*-Anstieg zurück, wie er über den *cut-off* Level im zweiten Testlauf erzielt wird. Auch für die Vergabe der Klassifikation erweist sich der Aufbau unterschiedlicher Trainingsmengen für die einzelnen Fachteilgebiete als positiv. Schließlich wirkt sich auch eine zunehmende Textlänge der Erschließungsgrundlage positiv auf das Indexierungsergebnis – im Sinne einer Übereinstimmung mit dem intellektuell generierten Vergleichsindexat – aus.

Mit Blick auf die praktische Umsetzung einer prozessunterstützenden automatischen Indexierung zeichnet sich aus der Zusammenschau der durchgeführten Testläufe ab, dass der Aufwand beim Aufbau fachteilgebietsspezifi-

[128] Für die Einschränkung der automatisch generierten Klassifikationsnotationen, so ging aus den zusätzlichen Auswertungsschritten des ersten Testlaufs hervor (siehe Anhang), ließe sich mit Abstandsmessungen der Konfidenzwerte arbeiten. Liegen die Konfidenzwerte der automatisch vorgeschlagenen Klassifikationsnotationen verhältnismäßig weit auseinander, lässt sich davon ausgehen, dass die nachfolgend vorgeschlagenen Klassifikationsnotationen in den meisten Fällen nicht intellektuell vergeben würden.

scher Versionen der Indexierungssoftware nicht im Verhältnis zum Qualitätsgewinn steht, den dieser mit sich bringen würde. Nicht nur ist die dadurch erzielte Erhöhung der Indexierungspräzision – gemessen an der intellektuellen Erschließung – geringer als bei der Festlegung eines *cut-off* Levels und einer gleichzeitigen Erhöhung des *Precision*-Wertes in den Systemeinstellungen, sondern auch die Vorselektion nach Fachteilgebieten und die Pflege derartiger Versionen der Indexierungssoftware ließen sich nur schwer ohne größeren Aufwand in den alltäglichen Geschäftsgang integrieren. Eine automatische Vor-Indexierung – in Anlehnung an das Schalenmodell nach Jürgen Krause – gezielt für die Erschließung der Randbereiche der Datenbank SOLIS einzusetzen, scheint somit nur schwer möglich.[129]

Zusätzlich geht aus der Studie hervor, dass für eine Evaluation der automatisch generierten Indexierungsergebnisse eine Unterscheidung der maschinell vergebenen Deskriptoren zwischen sachlichen Deskriptoren und Schlagwörtern aus den Sonderlisten sinnvoll ist. Sie dient einer differenzierteren Bewertung des automatisch generierten Indexierungsergebnisses. Auch hier zeichnet sich weiterer Forschungsbedarf ab. So ließe sich das Verhalten der Indexierungssoftware in Bezug auf bestimmte Erschließungsrichtlinien für die Sonderlisten näher untersuchen. Bisher lassen sich in diesem Zusammenhang zwei Problembereiche feststellen. Zum einen wird den speziellen Indexierungsregeln, wie sie für die sozialwissenschaftliche Fachinformation gelten, aus unterschiedlichen software- und verfahrenstechnischen Gründen nicht immer entsprochen. Teilweise fehlen Geographika oder das geographische Upposting wird nicht befolgt. Zum anderen führen diese Regeln unter Umständen dazu, dass inhaltlich fehlerhafte Indexierungsvorschläge generiert werden. Hierbei handelt es sich oftmals um sehr enge Verknüpfungen, da hierfür häufig eine größere Anzahl an Dokumenten in der Trainingsmenge existiert, für die diese Verknüpfung regelwerkskonform ist. Derartige regelwerksbegründete fehlerhafte Verknüpfungen ließen sich durch Regeldefinitionen einschränken. Hierzu ist es notwendig, auf der Indexierungsebene der einzelnen Dokumente bestimmte Verknüpfungen festzuschreiben bzw. auszu-

[129] Eine entsprechende Erhöhung der *Precision*, so zeigte sich bei den Vorbereitungen des vierten Testlaufs (siehe Kap. 3.3), war bei den reduzierten fachteilgebietsspezifischen Trainingsmengen gar nicht möglich.

schließen.[130] In diesem Zusammenhang erfordert etwa der Gebrauch bestimmter Geographika eine Regeldefinition, da durch das Upposting manche Deskriptoren zu häufig vergeben werden. Auch bei Länder- und Gebietsnamen, die unterschiedliche historische Bezeichnungen aufweisen (z.B. Ceylon/Sri Lanka, Palästina), erscheinen Regeldefinitionen sinnvoll. Gleiches gilt für historische Begriffe, wie Zeitangaben.

Schließlich könnte ein weiteres Forschungsfeld darin bestehen, noch stärker das Suchverhalten der Nutzer zu untersuchen. So ließe sich durch die Installation einer Tagging-Funktion Aufschluss darüber erhalten, welche Bezeichnungen Nutzer wählen, um Begriffe zu repräsentieren.

Reflexion des Evaluationsverfahrens

Nicht nur die inhaltliche Erschließung ist von Vagheit und Unschärfe begleitet – dies wird bei der Bewertung eines automatischen Erschließungsverfahrens anhand intellektuell generierter Vergleichsdaten gleichwohl besonders deutlich. Auch die Evaluation selbst wird ihrerseits von der eigenen Interpretation beeinflusst, insbesondere wenn sie auf der Ebene der einzelnen Dokumente vorgenommen wird. Diese und ähnliche Problematiken bei der Evaluation, sollen im Folgenden abschließend reflektiert werden.

Zunächst erschien die intellektuelle Indexierung – die Inkonsistenz der Erschließung zwischen unterschiedlichen Indexierern bereits berücksichtigt – nicht immer nachvollziehbar. So zeigten sich vereinzelt offenkundige Nachlässigkeiten bei der intellektuellen Indexierung. Dies ist bei der anfallenden Erschließungsmenge und der wenigen Zeit, die je dokumentarischer Bezugseinheit veranschlagt wird, verständlich. In einzelnen Fällen wurden etwa bestimmte Deskriptoren nicht vergeben, obgleich sie im Titel aufgeführt waren.[131] Auch bei der *Klassifikation* gab es Fälle, in denen es schwerfiel, die

[130] Als Beispiele lassen sich etwa die Verknüpfungen der Begriffe „Zusammenarbeit“ mit dem Deskriptor „Entwicklungsland“, „Indien“ mit dem Schlagwort „Hinduismus“ sowie „BRD“ und „DDR“ mit dem Deskriptor „Wiedervereinigung“ anführen.

[131] Ein Beispiel hierfür ist die Erschließung eines Dokuments mit dem Titel „Mehrebenen-Gouvernance in der Umweltpolitik und die Devolution Großbritanniens“. Der Deskriptor „Governance“ wurde nicht intellektuell, gleichwohl jedoch maschinell vergeben. Bei der Auswertung wurde die Vergabe des Titelbegriffs dennoch „nur“ als „zusätzlich passender Deskriptor“ gewertet.

intellektuelle Indexierung nachvollziehen.[132] Die Indexierungssoftware lieferte mitunter Ergebnisse, die einzig auf der Grundlage von Titel und Abstract bzw. Autorenreferat und Inhaltsverzeichnis den Inhalt der Publikation treffender wiederzugeben schienen.[133] Ebenso gab es Fälle, in denen Deskriptoren intellektuell offenkundig nicht regelkonform vergeben wurden.[134] Dies führt dazu, dass die Indexierungssoftware die Vergabe des entsprechenden Deskriptors fehlerhaft trainiert.

Einen besonderen Problembereich bei der Evaluation bildete die Bewertung der Deskriptoren und Klassifikationen, die vom intellektuellen Indexat abwichen. In ungenügender Weise ließ sich etwa die mitunter sehr enge begriffliche Überschneidung bei den intellektuell und automatisch vergebenen Deskriptoren abbilden. Aus dem Abgleich des automatisch generierten Indexats mit dem intellektuellen Erschließungsergebnis wurde gelegentlich deutlich, dass sich mit der Kombination unterschiedlicher Schlagwörter durchaus sehr ähnliche Inhalte abdecken ließen.[135] Des Weiteren ließ sich bei der Evaluation nur unbefriedigend wiedergeben, wenn die Indexierungssoftware mitunter nahezu das gesamte engere Begriffsumfeld eines Deskriptors, nicht allerdings genau den intellektuell vergebenen Deskriptor generierte.[136]

Wie bereits oben angerissen, ermöglichten zahlreiche maschinell generierte Schlagwörter zusätzliche Sucheinstiege. Auch dies ließ sich kaum bei der

---

[132] Exemplarisch hierfür ist die Vergabe der Klassifikation „Kultursoziologie, Kunstsoziologie“ bei einer Publikation mit dem Titel „Weltmacht Indien? Die Rolle Indiens in einer multipolaren Weltordnung“.

[133] Hier lässt sich beispielhaft ein automatisch generiertes Indexat zu einer Publikation über antisemitische Denkmuster anführen, in dem die Klassifikation „Sozialpsychologie“ vergeben wurde. Intellektuell wurden hingegen die beiden Klassifikationen „politische Willensbildung, politische Soziologie“ sowie „allgemeine Geschichte“ generiert.

[134] Dies betrifft etwa den Deskriptor „Transformation“. Entsprechend der *Scope Note* wird er „nur im Zusammenhang mit dem umfassenden Systemwechsel in den ehemaligen sozialistischen Ländern“ verwendet. Ansonsten gelten die Schlagwörter „ökonomischer – “, „politischer –“ oder „sozialer Wandel“ oder das Schlagwort „sozioökonomische Entwicklung“ (IZ 2002: 290).

[135] Eingebettet in ein semantisches Umfeld aus den Deskriptoren „historische Entwicklung“ und „politische Entwicklung“ ließe sich etwa diskutieren, inwiefern die Deskriptoren „institutioneller Wandel“ – wie er intellektuell generiert – und „Institution“ sowie „politische Institution“ – wie sie maschinell vergeben wurden – nicht sehr nah zusammenliegen. Da die entsprechenden maschinell zugeordneten Bezeichnungen keinen Informationsmehrwert bieten, wurden sie auch nicht als „zusätzlich passende Deskriptoren“ eingestuft.

[136] Exemplarisch hierfür ist ein Indexat, bei dem zwar die Deskriptoren „Krieg“ und „Gewalt“ sowie „Konflikt“ und „politischer Konflikt“ nicht allerdings das zusätzlich intellektuell vergebene Schlagwort „militärischer Konflikt“ generiert wurde.

Evaluation berücksichtigen. Dazu gehörten sowohl sehr spezifische als auch eher allgemein gehaltene Deskriptoren.[137] Eine entsprechende Suchstrategie würde etwa eher einem ungeschulten Nutzer entsprechen. Ebenso gehörten zu den zusätzlich vergebenen Deskriptoren Formschlagwörter, wie „Lehrbuch" oder „Einführung", die durch entsprechende Formulierungen in Titel und Inhaltstext generiert wurden. Auch dies kann für das Information Retrieval von Vorteil sein und sich gerade für neue Nutzerkreise als eine Hilfe bei der Recherche erweisen. Zusätzlich zur Klassifikation ließe sich hierdurch nach einführender Literatur suchen.[138] Daneben wurden in manchen Fällen etwa bestimmte Ländernamen mit angegeben, ohne dass sie im Abstract vorkamen.[139] Mitunter konnte davon ausgegangen werden, dass diese auch in der Publikation vorkamen, doch war unklar, wie bei der Bewertung des Indexats damit verfahren werden sollte.

Bei der Klassifikation ließen sich zahlreiche Publikationen mehreren Disziplinen zuordnen. Dies wird dadurch begünstigt, dass manche Klassifikationsnotationen thematisch sehr nah beieinander liegen.[140] Nur wenige Klassifikationsvorschläge ließen sich oftmals daher als eindeutig fehlerhaft bewerten. Tendenziell entspricht dieses Indexierungsverhalten auch einer weiter gefassten Auslegung der Erschließungsrichtlinie für die Klassifikation (vgl. GESIS 2005). Klassifikationsvorschläge, die sich nicht mit der intellektuellen Klassi-

[137] Für Letztere lassen sich als Beispiel die Deskriptoren „Macht" und „Herrschaft" anstelle von „politischer Macht" und „politischer Herrschaft" sowie „Frau" anstelle von „Frauenfrage" und „Geschlecht" anführen.

[138] Unter Umständen wurden durch die Indexierungssoftware auch Begriffe vergeben, die in den Bereich der Körperschaften fielen. So wurde für ein Dokument etwa der Deskriptor „Arendt, H." vergeben, da es sich um einen Betrag im Rahmen eines Hannah-Arendt-Kolloquiums handelte. Auch dies kann für bestimmte Suchanfragen eine Unterstützung darstellen. Noch einmal sei in diesem Zusammenhang erwähnt, dass für die Suche in der Datenbank SOLIS über *sowiport* ein Vorschlagsdienst zum Auffinden weiterführender Suchbegriffe aus dem Thesaurus Sozialwissenschaften vorliegt (vgl. Mayr et al. 2009). Dieser ermöglicht die sinnvolle Eingrenzung größerer Treffermengen. Gleichzeitig gilt es, sich gleichwohl bewusst zu machen, dass die Präzision der Suchergebnisse sehr stark darunter leiden kann.

[139] Exemplarisch hierfür ist eine Publikation mit dem Titel „Lebensbedingungen und Wohlbefinden in Europa". Obgleich im Autorenreferat keinerlei Ländernamen aufgeführt waren, wurden von der Indexierungssoftware aufgrund der Kontexterkennung ein (Testlauf I), sieben (Testlauf III) bzw. zwei (Testlauf IV) Geographika vergeben.

[140] Hier lassen sich neben „Religionssoziologie" oder „Philosophie, Ethik, Religion" die Klassifikationen „Friedens- und Konfliktforschung" und „internationale Beziehungen" sowie „Staat, politisches System" und „politische Willensbildung, politische Soziologie, politische Kultur" anführen, die intellektuell und – auf dieser Trainingsgrundlage beruhend – vielfach auch maschinell gemeinsam vergeben werden. Bei der automatischen Klassifikation lässt sich dies mitunter daran belegen, dass die Konfidenzwerte dieser Klassifikationsvorschläge, wenn sie gemeinsam vergeben wurden, sehr nah beieinander lagen.

fikation deckten, ließen sich allerdings häufig nur einzelnen Ausschnitten der Publikation zuordnen. Dies spiegelt sich auch darin wider, dass sich nur wenige Klassifikationsvorschläge als eindeutig ‘zusätzlich passende Klassifikation’ charakterisieren ließen. Daneben konnte es im Ermessen des Indexierers liegen, welche Klassifikationsnotation er bei der intellektuellen Erschließung auswählte.[141] Schließlich wurde die Bewertung der Klassifikation sehr stark dadurch erschwert, dass die Publikationen nicht in Autopsie vorlagen.

Diese Reflexion des Evaluationsverfahrens verdeutlicht die relative Aussagekraft der vorliegenden Arbeit. Bei der Evaluation bis auf die Dokumentebene vorzudringen, schließt in besonderer Weise einen interpretativen Prozess mit ein. In dieser Hinsicht sind sich inhaltliche Repräsentation und ihre Evaluation sehr ähnlich. Gleichwohl sollte aus den erzielten Untersuchungsergebnissen ein Beitrag hervorgegangen sein, der zu weiteren Tests, wie etwa umfangreichen Retrieval-Tests, anregt, um eine bereits aktuell mögliche semiautomatische Indexierung in der sozialwissenschaftlichen Fachinformation zu etablieren und auf längere Sicht hin zu optimieren.

[141] Als eine Frage des eigenen Ermessens erschien des Öfteren auch die Evaluation des Indexierungsergebnisses. In den meisten Fällen ging es dabei um die fachlichen Grenzen der unterschiedlichen Wissenschaftsgebiete und -disziplinen.

Literaturverzeichnis

Bertram, Jutta (2005) Einführung in die inhaltliche Erschließung. Grundlagen – Methoden – Instrumente. Würzburg: ERGON-Verlag.

Castells, Manuel (2001) Der Aufstieg der Netzwerkgesellschaft. Opladen: Leske + Budrich.

David, Claire/Giroux, Luc/Bertrand-Gastaldy, Suzanne/Lanteigne, Diane (1995) Indexing as a problem solving – a cognitive approach to consistency. In: Proceedings of the American Society for Information Science. Jg. 32, 49-55 http://www.cais-acsi.ca/proceedings/1995/david_1995.pdf Zugriff am 30.04.2011.

Deerwester, Scott/Landauer, Thomas/Furnas, George/Dumais, Susan T./Harshman, Richard (1990) Indexing by Latent Semantic Analysis, in: Journal of the American Society for Information Science, vol. 41, 391-407. http://lsi.research.telcordia.com/lsi/papers/JASIS90.pdf. Zugriff am 09.04.2011.

Deutsche Nationalbibliothek (DNB) (2010) PETRUS. Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek. Interne Präsentation, Mai 2010.

Ferber, Reginald (2003) Information-Retrieval – Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web. Heidelberg: dpunkt.verlag. http://information-retrieval.de/irb/gifs/schema-ifs-vrm.gif Zugriff am 09.04.2011.

Fuhr, Norbert (2004[5]) Theorie des Information Retrieval I: Modelle, in: Kuhlen, Rainer et al. (Hg.) Grundlagen der praktischen Information und Dokumentation, München: Saur, 207-214.

Furnas, G.W./Landauer, T.K. (1987) The Vocabulary Problem in Human-System Communication: an Analysis and a Solution, 32(1), 4-8, in: communications of the ACM, 30(11), 964-971.

Gerards, Michael/Gerards, Andreas/Weiland, Peter (2006) Der Einsatz der automatischen Indexierungssoftware AUTINDEX im Zentrum für Psychologische Information und Dokumentation (ZPID) http://www.zpid.de/download/PSYNDEXmaterial/autindex.pdf Zugriff am 02.04.2011.

Gerards, Michael (2011) Semiautomatische Erschließung von Psychologie-Information. Präsentation im Rahmen des PETRUS-Workshops an der Deutschen Nationalbibliothek Frankfurt/Main, 21./22. März 2011 http://files.d-nb.de/pdf/petrus/semiautomatische_erschliessung_zpid.pdf Zugriff am 22.05.2011.

GESIS │Informationszentrum Sozialwissenschaften (Hg.) (2005) Regelwerk für die Literaturdokumentation Sozialwissenschaften. Regeln für die inhaltliche Erschließung sozialwissenschaftlicher Literatur.

Gödert, Winfried/Lepsky, Klaus (1997) Semantische Umfeldsuche im Information Retrieval in Online – Katalogen, in: Kölner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft, 7.

Groß, Thomas (2010) Die Implementierung eines automatischen Indexierungsverfahrens am Beispiel der Deutschen Zentralbibliothek für Wirtschaftswissenschaften. Masterarbeit im Rahmen des postgradualen Fernstudiums Master of Arts. http://edoc.hu-berlin.de/master/gross-thomas-2010-05-08/PDF/gross.pdf Zugriff am 10.11.2010.

Groß, Thomas/Faden, Manfred (2010) Automatische Indexierung elektronischer Dokumente an der Deutschen Zentralbibliothek für Wirtschaftswissenschaften. In: Bibliotheksdienst, 44. Jg., 12, 1120-1135.

Holl, Daniela (2009) Search Term Recommender auf Basis der Software MindServer. Diplomarbeit im Studiengang Computervisualistik, Universität Koblenz-Landau. Koblenz.

Hooper, R.S. (1965) Indexer consistency tests – Origin, measurements, results and utilization. Bethesda: IBM.

Hübner, Juliane/Groth, Elke (2004) Darstellung der Vor- und Nachteile ausgewählter automatischer Indexierungsverfahren im Vergleich zum intellektuellen Indexieren. Magisterarbeit, Humboldt-Universität zu Berlin: Institut für Bibliotheks- und Informationswissenschaft. Berlin. Kontakt: hueju@cmb.hu-berlin.de

Ingwersen, Peter (1996) The cognitive framework for information retrieval: A paradigmatic perspective. In: Krause, Jürgen/Herfurth, Matthias/Marx, Jutta (Hg.) Herausforderungen an die Informationswissenschaft. Informationsverdichtung. Informationsbewertung und Datenvisualisierung. Proceedings des 5. Internationalen Symposiums für Informationswissenschaft (ISI), Humboldt-Universität zu Berlin, 17.-19. Oktober 1996.

Institut für Technologiefolgen-Abschätzung (ITA) der Österreichischen Akademie der Wissenschaften (Hg.) (2010) Google, Google Scholar und Google Books in der Wissenschaft. Steckbrief 3 im Rahmen des Projekts Interactive Science.http://epub.oeaw.ac.at/ita/ita-projektberichte/d2-2a52-3.pdf Zugriff am 10.05.2011.

Informationszentrum Sozialwissenschaften (IZ) (2002) Thesaurus Sozialwissenschaften. Alphabetischer Teil/Systematischr Teil. Bonn: Informationszentrum Sozialwissenschaften.

Informationszentrum Sozialwissenschaften (IZ) (2006) Thesaurus Sozialwissenschaften. Alphabetischer Teil/Systematischr Teil. Bonn: Informationszentrum Sozialwissenschaften.

KASCADE – http://www.uni-duesseldorf.de/projekte/kascade/kas_home Zugriff am 06.04.2011.

Keil, Stefan/Tiesler, Philipp et al. (2010) Automatische Erschließung für die Datenbank SOLIS. Internes Arbeitspapier von GESIS │Leibniz-Institut für Sozialwissenschaften.

Krause, Jürgen (1996) Informationserschließung und –bereitstellung zwischen Deregulation, Kommerzialisierung und weltweiter Vernetzung – Schalenmodell. IZ-Arbeitsbericht Nr. 6, Bonn: Informationszentrum Sozialwissenschaften.

Krause, Jürgen (2006) Shell Model, Semantic Web and Web Information Retrieval. In: Harms, Ilse/Luckhardt, Heinz-Dirk/Giessen, Hans W. (Hg.) Information und Sprache. Beiträge zu Informationswissenschaft, Computerlinguistik, Bibliothekswesen und verwandten Fächern. Festschrift für Harald H. Zimmermann. K.G. Saur: München, 95-106.

Kuhlen, Rainer (2004) Informationsaufbereitung III: Referieren (Abstracts – Abstracting – Grundlagen). In: Ders./Seeger, Thomas/Strauch, Dieter (Hg.) Grundlagen der praktischen Information und Dokumentation, 5. völ. neu gef. Auflage. München: K.G. Saur, 189-206.

Lancaster, Frederick, W. (19982) Indexing and abstracting in theory and practice. London: Library Association Publishing (Kap. 2, 3, 5-6 u. 11).

Landauer, Thomas/Foltz, Peter W./Laham, Darrell (1998) Introduction to Latent Semantic Analysis, in: Discourse Processes, 25, 259-284.

Leibniz-Gemeinschaft – Der Senat (Hg.) (2005) Stellungnahme zur Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen (GESIS). http://www.wgl.de/download.php?fileid=61 Zugriff am 01.05.2011.

Leininger, K. (2000)Interindexer consistency in PsycINFO, in: Journal of Librarianship and Information Science. 32: 4-8)

Lingelbach-Hupfauer, Carmen/Laute, Hartwig (2009) Die semiautomatische Indexierung von Zeitungsartikeln. In: Info 7, Jg. 24, Heft 2, 48-50.

Lingelbach-Hupfauer, Carmen (2011) Die semi-automatische Erschließung von Zeitungs- und Zeitschriftenartikeln in der Pressedokumentation des ZDF. Präsentation im Rahmen des PETRUS-Workshops an der Deutschen Nationalbibliothek Frankfurt/Main, 21./22. März 2011 http://files.d-nb.de/pdf/petrus/semi-automatische_indexierung_zdf.pdf Zugriff am 19.05.2011.

Mayr, Philipp (2010) Information Retrieval – Mehrwertdienste für Digitale Bibliotheken. Crosskonkordanzen und Bradfordizing. Bonn: GESIS │Leibniz-Institut für Sozialwissenschaften.

Mayr, Philipp/Mutschke, Peter/Schaer, Philipp/Sure, York (2009) Mehrwertdienste für das Information Retrieval: das Projekt IRM. http://www.ib.hu-berlin.de/~mayr/arbeiten/IRM-ISKO09.pdf Zugriff am 03.04.2011.

MILOS – http://www.ub.uni-duesseldorf.de/projekte/milos/mil_home Zugriff am 05.04.2001

Mittelbach, Jens/Probst, Michaela (2006) Möglichkeiten und Grenzen maschineller Indexierung in der Sacherschließung – Strategien für das Bibliothekssystem der Freien Universität Berlin. In: Berliner Hand-reichungen zur Bibliotheks- und Informationswissenschaft, Heft 183. Berlin.

http://webdoc.sub.gwdg.de/ebook/serien/aw/Berliner_Handreichungen/h183.pdf Zugriff am 03.04.2011.

Nohr, Holger (2004[5]) Theorie des Information Retrieval II: Automatische Indexierung, in: Kuhlen, Rainer et al. (Hg.) Grundlagen der praktischen Information und Dokumentation, München: Saur, 215-225.

Nohr, Holger (2005[3]) Grundlagen der automatischen Indexierung – Ein Lehrbuch. Berlin: Logos-Verlag.

Organisation der Vereinten Nationen für Erziehung, Wissenschaft und Kultur – Büro für internationale Normen und Rechtsfragen (Hg.) (1979) Empfehlungen zur internationalen Vereinheitlichung der Statistiken über Wissenschaft und Technologie der UNESCO-Generalkonferenz vom 28.11.1978.

Puzicha, Jan (2009) Informtionen finden! Intelligente Suchmaschinentechnologie & automatische Kategorisierung. Technical Whitepaper – Grundlagen der Informationsgewinnung. MindServer. Publikation der Firma Recommind.

Recommind (2011) Automatische Kategorisierung http://www.recommind.de/loesungen/Automatische_Kategorisierung Zugriff am 30. März 2011.

Rolling, L. (1981) Indexing consistency, quality and efficiency. In: Information Processing & Management, 17. Jg., 69-76.

Saarti, Jarmo (2002) Consistency of subject indexing of novels by public library professionals and patrons. In: Journal of Documentation, Jg. 58, 1, 49-65.

Salton, Gerard (1987) Information Retrieval – Grundlegendes für Informationswissenschaftler. Hamburg: McGraw-Hill.

Schöning-Walter, Christa (2011) Automatische Erschließungsverfahren für Netzpublikationen. Stand der Arbeiten im Projekt PETRUS. Präsentation bei einem Kolloquium von GESIS │Leibniz-Institut für Sozialwissenschaften, Februar 2011.

Siegmüller, Renate (2007) Verfahren der automatischen Indexierung in bibliotheksbezogenen Anwendungen. In: Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft, Heft 214. Berlin. http://www.ib.hu-berlin.de/~kumlau/handreichungen/h214/h214.pdf Zugriff am 03.04.2011.

Sparck Jones, Karen (1973) Automatic Indexing. In: Journal of Documentation, 30, 393-432.

Stahl, Matthias/Binder, Giesbert/Riege, Udo (2005) Befragung des akademischen Mittelbaus im Fach Soziologie, IZ-Arbeitsbericht Nr. 34.

Steinbicher, Jochen (2001)Zur Theorie der Informationsgesellschaft. Ein Vergleich der Ansätze von Peter Ducker, Daniel Bell und Manuel Castells. Opladen: Leske + Budrich.

Stock, Wolfgang G. (2007) Information Retrieval. Informationen suchen und finden. München: Oldenbourg Verlag.

Stock, Wolfgang G./Stock, Mechtild (2008) Wissensrepräsentation. Informationen auswerten und bereitstellen. München: Oldenbourg Verlag.

Support Vector Machine – Wikipedia http://de.wikipedia.org/wiki/Support_Vec-tor_Machine Zugriff am 09.04.2011.

Wersig, Gernot (1978) Thesaurus-Leitfaden. Eine Einführung in das Thesaurus-Prinzip in Theorie und Praxis. München: Verl. Dokumentation.

Wissel, Verena (2011) Erfahrungsbericht und Schlussfolgerungen des DIPF zur Erprobung automatischer Erschließung. Präsentation im Rahmen des PETRUS-Workshops an der Deutschen Nationalbibliothek Frankfurt/Main, 21./22. März 2011. http://files.d-nb.de/pdf/petrus/dipf_erfahrungsbericht.pdf Zugriff am 15.05.2011.

Womser-Hacker, Christa (2004[5]) Theorie des Information Retrieval III: Evaluierung, in: Kuhlen, Rainer/Seeger, Thomas/Strauch, Dieter (Hg.) Grundlagen der praktischen Information und Dokumentation, 5. völ. neu gef. Auflage. München: K.G. Saur, 227-235.

Xu, Chen (2007) Indexing Consistency between Online Catalogues. Dissertation, Humboldt-Universität zu Berlin, Institut für Bibliotheks- und Informationswissenschaft. Berlin. http://edoc.hu-berlin.de/dissertationen/chen-xu-2008-05-14/PDF/chen.pdf Zugriff am 03.04.2011.

Anhang

Übersicht der Begriffsbeziehungen im Thesaurus Sozialwissenschaften

	Kürzel (Englisch)	**Langform**	**Beispiel**
Äquivalenz-relation	USE [*use*]	Benutze Synonym	Migrationssteuerung USE Migrationspolitik
		Benutze Kombination	Migrationsdruck USE Migration + Ursache
	UF [*use(d) for*]	Benutzt für Synonym	Migration UF räumliche Mobilität
Hierarchie-relation	BT [*broader term*]	Oberbegriff	Migrationspolitik BT Bevölkerungspolitik
	NT [*narrower term*]	Unterbegriff	Migrationspolitik NT Asylpolitik
Assoziations-relation	RT [*related term*]	Verwandter Begriff	Migration RT Ortsbezogenheit

Tabelle 10: Begriffsbeziehungen im Thesaurus Sozialwissenschaften.

Gesamtdarstellung der MindServer-Einstellungen für die Testläufe I bis IV

	Testlauf 1 Gesamtstichprobe		**Testlauf 2 Gesamtstichprobe**		**Testlauf 3 Fachteilgebiete** **Soziologie, Politik- und Geisteswissenschaften**		**Testlauf 4 Fachteilgebiete** **Soziologie, Politik- und Geisteswissenschaften**	
	Thesaurus	Klassifikation	Thesaurus	Klassifikation	Thesaurus	Klassifikation	Thesaurus	Klassifikation
Minimum Trainingsdokumente pro Kategorie	1	1	**20**	**20**	1	1	**20**	**20**
Maximum Trainings-dokumente pro Kategorie	25.0000	25.000	25.000	25.000	25.000	25.000	25.000	25.000
Minimum Deskripto-ren/Klass. pro Dokument	keine Ein-schränkung (-1)	-1	-1	-1	-1	-1	29	**2**
Maximum Deskripto-ren/Klass. pro Dokument	keine Ein-schränkung (-1)	-1	**10**	**5**	-1	-1	30	**5**
Verhältnis Recall/Precision	20.0	20.0	**-20.0**	20.0	20.0	20.0	**0**	20.0
Minimale Textlänge	keine Ein-schränkung (-1)	-1	**25**	**25**	-1	-1	25	25

Tabelle 11: Systemeinstellungen der Indexierungssoftware MindServer Testläufe I bis IV.

Zusätzliche Auswertungsschritte des ersten Testlaufs

Im Anschluss an die allgemeine Auswertung des ersten Testlaufs erfolgte eine Reihe weiterer Auswertungsschritte. Sie bildeten eine erste Annäherung an das Untersuchungsfeld und dienten dazu, ein differenzierteres Bild über die Qualität der Indexierungsergebnisse zu erhalten. Aufgrund des engen zeitlichen Rahmens bezogen sich diese zusätzlichen Auswertungsschritte lediglich auf Unterstichproben.

Als Einstieg wurden die Datensätze der Gesamtstichprobe in Gruppen unterteilt. Die Gruppen setzten sich folgendermaßen zusammen:

- Datensätze (Anzahl 40) mit 30 und mehr maschinell generierten Deskriptoren,
- Datensätze (Anzahl 14) mit dreimal so vielen maschinell wie intellektuell vergebenen Deskriptoren,
- Datensätze (Anzahl 38), bei denen mehr als die Hälfte der intellektuell vergebenen Deskriptoren auch maschinell vergeben wurde,
- Datensätze (Anzahl 96), bei denen insgesamt weniger Deskriptoren maschinell als intellektuell vergeben wurden.

Auf dieser Grundlage wurden Hypothesen zum Vorgehen der Indexierungssoftware aufgestellt.[142] Da die bis dato erfolgte Auswertung der Indexierungsergebnisse (siehe oben) nicht (mehr) auf der Ebene der einzelnen Datensätze in Form von aDIS- sowie Ausdrucken der automatisch erstellen Indexierungsergebnisses vorlag, erwies sich dieses Vorgehen bald als zu zeitaufwändig und zu wenig ertragreich. Bis eine nennenswerte Anzahl von Datensätzen der Gesamtstichprobe durchgesehen werden konnte, um eine Hypothese zu verifizieren oder falsifizieren, nahm zu viel Zeit in Anspruch. Gleichwohl ergab sich hieraus ein Gefühl für die Qualität der automatisch generierten Indexierungsergebnisse. Erste Ergebnisse über Auffälligkeiten der Indexierungssoftware bei der Erschließung fanden Eingang in eine Dokumentation.

[142] Eine Hypothese lautete etwa: Die Anzahl der automatisch generierten Deskriptoren ist abhängig von der Abstract-Art (Abstract/Autorenreferat).

Nachfolgend wurden aDIS sowie Mindserver-Abfragen zu denjenigen Datensätzen durchgeführt, die die höchsten Überschneidungsmengen mit der intellektuellen Erschließung aufwiesen. Unter Verwendung der Ergebnistabelle, wie sie bis dato von den GESIS-Mitarbeiter-/innen erstellt worden war, wurden dafür folgende Kriterien, die gleichzeitig erfüllt sein mussten, aufgestellt:

- eine Überschneidungsmenge von mindestens vier Deskriptoren,
- keine fehlenden Titelbegriffe,
- keine „falschen" Deskriptoren,
- Hauptklassifikation gefunden.

Vor dem Hintergrund, dass in anderen Datenbanken sowie in Bibliothekskatalogen intellektuell häufig deutlich weniger Deskriptoren vergeben werden als in der Fachinformation Sozialwissenschaften – in der ZBW werden beispielsweise im Durchschnitt lediglich fünf Deskriptoren vergeben –, wurden für diejenigen automatisch generierten Indexate, die der intellektuellen Erschließung am nächsten kamen, verschiedene *cut-off* Levels getestet. Gemessen wurden die Überschneidungsmengen für die gemäß dem Konfidenzwert ersten fünf, sieben und zehn automatisch generierten Deskriptoren. Den Hintergrund dieses Vorgehens bildete das Ziel, durch ein geeignetes Maximum an Deskriptoren pro Indexat die Präzision der Indexierungsergebnisse zu erhöhen.

Es zeigte sich, dass diejenigen Deskriptoren, die auch intellektuell vergeben worden waren, häufig relativ hohe Konfidenzwerte erzielten. Da bei der Auswertung des ersten Testlaufs eine durchschnittliche Überschneidungsmenge von sechs Deskriptoren ermittelt worden war, wurden vor allem Datensätze mit fünf (Anzahl 44), sechs (Anzahl 31) und sieben (Anzahl 36) Überschneidungen ausgewertet.

Parallel dazu wurde die ursprüngliche Auswertungstabelle für den ersten Testlauf unter Berücksichtigung der Dokumentarten neu geordnet und zunehmend

weitere Variablen bei der Bewertung der Indexierungsergebnisse eingeführt. Im Einzelnen wurde jeder Datensatz nach folgenden Kriterien ausgewertet: [143]

- Dokumentart,
- Sprache des Dokuments,
- Abstract-Art (Abstract/Autorenreferat inkl. Sprache),
- Zeilenanzahl der Materialgrundlage in der aDIS-Anzeige,
- Datenbasis (z.B. IZ-SOZ oder PROGRIS),
- Anzahl der Deskriptoren (MindServer),
- Anzahl der Deskriptoren (intellektuell),
- Anzahl der Überschneidungen. An dieser Stelle wurde eine Differenzierung anhand der verschiedenen *cut-off* Levels durchgeführt. Jeweils in Klammern wurde angegeben, wie viele Überschneidungen unter den ersten fünf, sieben und zehn automatisch generierten Deskriptoren gezählt wurden.
- Anteil der Überschneidungen an der Ergebnismenge des MindServer (*Precision*-Wert),
- Anteil der Überschneidungen an der Ergebnismenge der intellektuellen Erschließung (*Recall*-Wert),
- Anteil der Überschneidungen an der Ergebnismenge (intellektuell) bei einem *cut-off* Level von zehn zulässigen Deskriptoren
- Bewertung der ersten fünf von MindServer vergebenen Deskriptoren anhand einer Skala von 1 bis 4 (inklusive der Zwischenwerte).[144] Zusätzlich wurde für die in dieser Form bewerteten ersten fünf automatisch generierten Deskriptoren aufgeführt, ob es sich um Deskriptoren

[143] Um ein vollständiges Bild der Analyse in diesem Stadium wiederzugeben, werden auch diejenigen Auswertungsvariablen aufgezählt, die bereits in der ursprünglichen Bewertungstabelle aufgeführt wurden.

[144] „1“ bedeutet „sehr gut“. D.h., der Deskriptor wurde auch intellektuell vergeben oder stellt einen zusätzlich von MindServer vergebenen passenden Deskriptor dar. „2“ steht für einen gut verwendbaren Deskriptor. Der Deskriptor wurde nicht intellektuell vergeben, bietet allerdings einen Informationsmehrwert. „3“ steht für einen Deskriptor der keinen Informationsmehrwert bietet. Meist sind die derart bewerteten Deskriptoren zu allgemein gehalten. „4“ steht für einen inhaltlich nicht zutreffenden oder irreführenden Begriff.

aus den Sonderlisten handelte.[145] In Klammern wurde in dieser Spalte gleichzeitig berücksichtigt, inwieweit unter den ersten fünf, sieben und zehn Deskriptoren weiterhin bestimmte Titelbegriffe fehlten.[146]

- Allgemeine Bewertung der inhaltlichen Wiedergabe des Dokuments unter Berücksichtigung der ersten fünf, sieben und zehn Deskriptoren anhand einer Skala aus „+“ (sehr gut/gut), „0“ (wenig aussagekräftig) und „–“ (mangelhaft).
- Auflistung, an welcher Stelle des Rankings entsprechend des Konfidenz-wertes sich der fünfte Deskriptor befindet, der sowohl intellektuell als auch von MindServer vergeben wurde (z.B. 15. Stelle).
- Wiedergabe der Konfidenzwerte der ersten beiden inhaltlichen Deskriptoren (z.B. 092; 0,79).

Hieran folgte die Bewertung der Klassifikationen:

- Anzahl der von MindServer vergebenen Klassifikationen,
- Anzahl der intellektuell vergebenen Klassifikationen,
- Auflistung der einzelnen intellektuell vergebenen Klassifikationen (dieses Feld war für eine im Anschluss vorgenommene „Vermessung“ der Gesamtstichprobe (280 Dokumente) anhand der intellektuell generierten Hauptklassifikationen erforderlich.
- Anzahl der Überschneidungsmenge bei der Klassifikation,
- Bewertung der maschinell vergebenen Klassifikationen anhand einer Skala von 1+ bis 3 bei einem *cut-off* Level von drei,[147]

---

[145] Im Einzelnen steht „(allg.)“ für Allgemeinbegriffe (z.B. Analyse oder Prozess), „(hist.)“ steht für historische Begriffe (z.B. Deutsches Reich oder 19. Jahrhundert), „(geo.)“ bzw. „(LRN)“ steht für geographische Deskriptoren (z.B. „alte Bundesländer“, „Entwicklungsland“ oder „Bundesrepublik Deutschland“). Eine Bewertung der ersten fünf Deskriptoren sah daher beispielsweise folgendermaßen aus: 1(geo), 1(geo), 1(geo), 2-3, 4(hist.).

[146] Für die obige exemplarische Darstellung einer Bewertung der ersten fünf Deskriptoren sah dies etwa folgendermaßen aus: 1(geo), 1(geo), 1(geo), 2-3, 4(hist.)(1)(1)(0). Während nach fünf bzw. sieben Deskriptoren noch jeweils ein Titelbegriff fehlte, war dies nach den ersten zehn von MindServer vergebenen Deskriptoren nicht mehr der Fall.

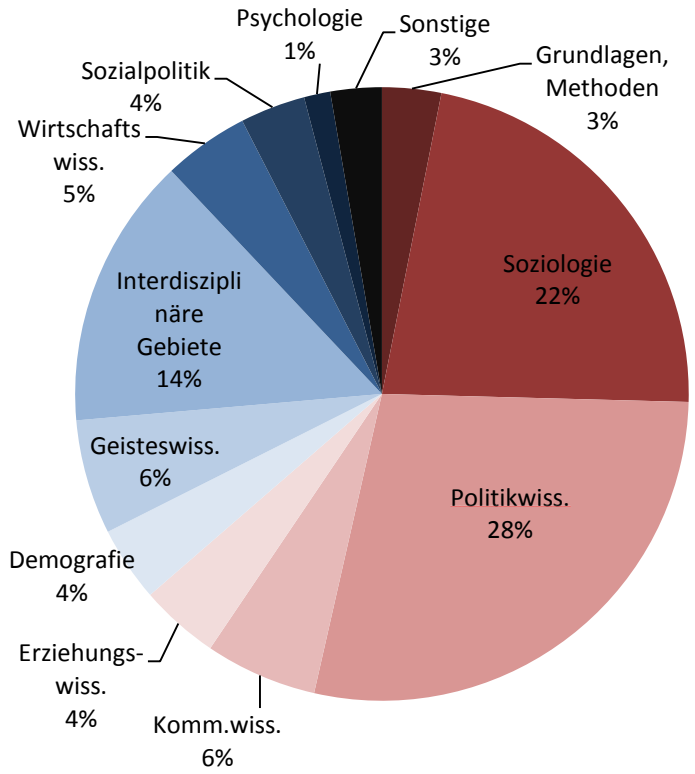
[147] „1+“ steht für die Hauptklassifikation. „1“ steht für eine Klassifikation, die auch intellektuell vergeben wurde. „2“ steht für eine automatisch vergebene Klassifikation, die durchaus auch intellektuell vergeben worden sein könnte. „3“ steht für eine inhaltlich unzutreffende bzw.

- Allgemeine Bewertung des Klassifikationsergebnisses erneut anhand einer Skala aus „+“, „0“ und „–“ bei einem *cut-off* Level von drei.
- Nennung, ob die Hauptklassifikation gefunden wurde oder nicht (ja/nein),
- Darstellung, an welcher Position des Rankings gemäß dem Konfidenzwert sich die Hauptklassifikation befindet (z.B. 2 oder 4),
- Auflistung der Anzahl „falscher“ von MindServer vergebener Klassifikationen,
- Wiedergabe, inwieweit von MindServer zusätzlich passende Klassifkationen gefunden wurden,
- Darstellung, bei welchem Konfidenzwert die ersten beiden von MindServer vergebenen Klassifikationen beginnen (z.B. 0,64; 0,27). Hieran lässt sich ablesen, wie „sicher“ sich die Indexierungssoftware bei der Vergabe der Klassifikationen ist bzw. inwieweit die Indexierungssoftware auf der Grundlage der Trainingsmenge auf ähnliche Dokumente zurückgreifen kann.

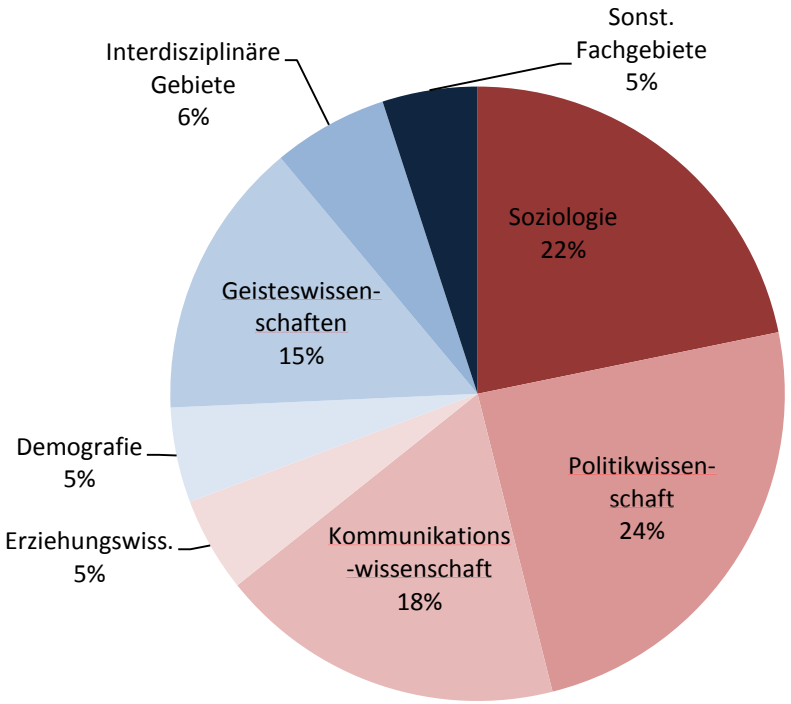
irreführende Klassifikation. Zwischenwerte wurden ebenfalls zugelassen. Eine exemplarische Bewertung sah somit folgendermaßen aus: 1, 3, 1+. Die Hauptklassifikation erscheint nach dem Konfidenzwert an dritter Stelle.

Verteilung der Fachteilgebiete

Verteilung der Fachteilgebiete in SOLIS



Verteilung der Fachgebiete in der Gesamtstichprobe (n=271)



Fachteilgebietsspezifische Voruntersuchung

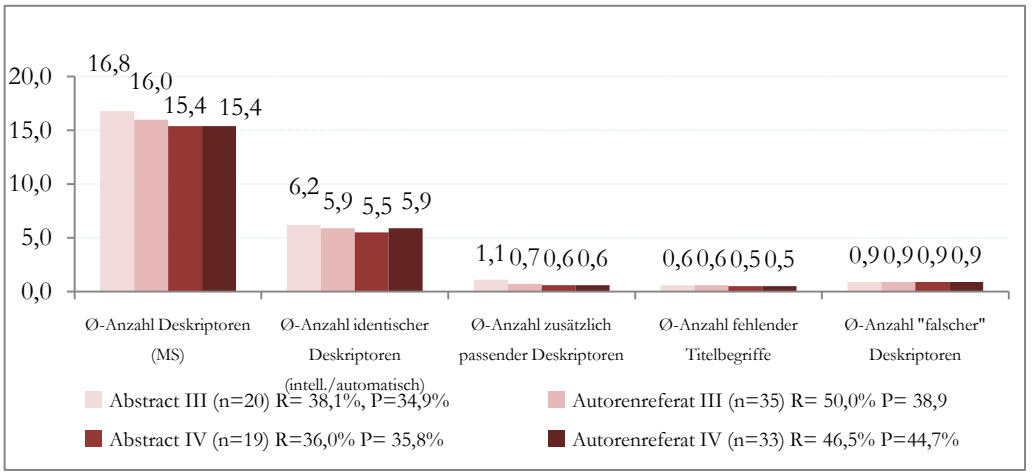
	Testlauf I (keine Einschränkung der Deskriptorenanzahl)			**Testlauf II (max. 10 Deskriptoren zulässig)**		
	Ø-Anzahl Überschn.	mind. 5 Über-schn.(%)	Haupt-klass. gef. (%) (1.Rang)	Ø-Anzahl Überschn.	mind. 5 Überschn. SW (%)	Haupt-klass. gef. (%) (1.Rang)
Soziologie n=56 (I) n=54 (II)	6,2	72,9	76,3 (42,4)	4,9	52,8	75,5 (47,2)
Politolo-gie n=68 (I) n=67 (II)	7,2	63,2	91,2 (52,9)	5,0	60,9	90,6 (60,9)
Pädagogik n=14 (I,II)	5,3	71,4	64,3 (21,4)	2,8	18,2	63,6

Tabelle 12: Exemplarische Gegenüberstellung von Kern- und Randbereichen der Datenbank SOLIS Testläufe I und II

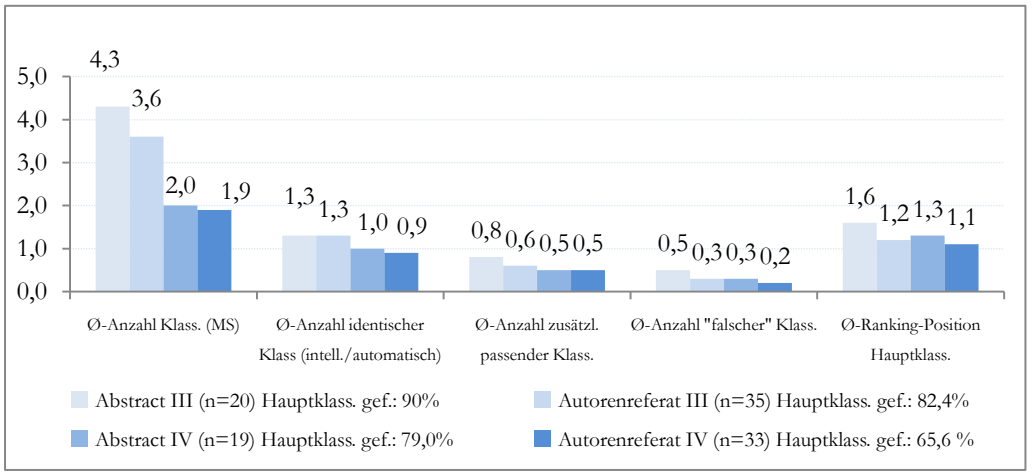
Vergleich der autom. generierten Indexierungsergebnisse nach Abstract-Art für die Fachteilgebiete Soziologie, Politik- und Geisteswissenschaften

Soziologie

Vergabe der Deskriptoren

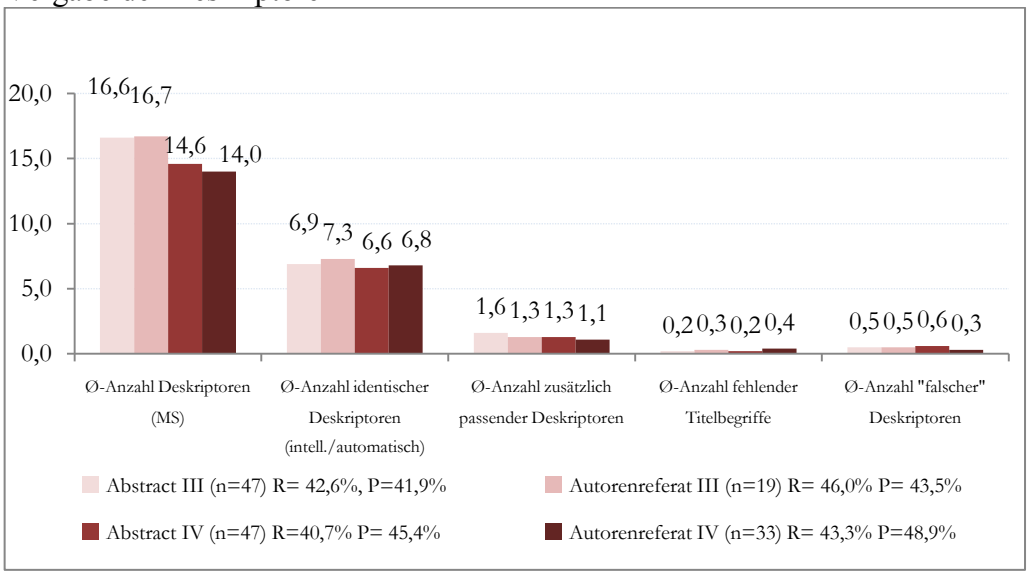


Zuordnung der Klassifikation

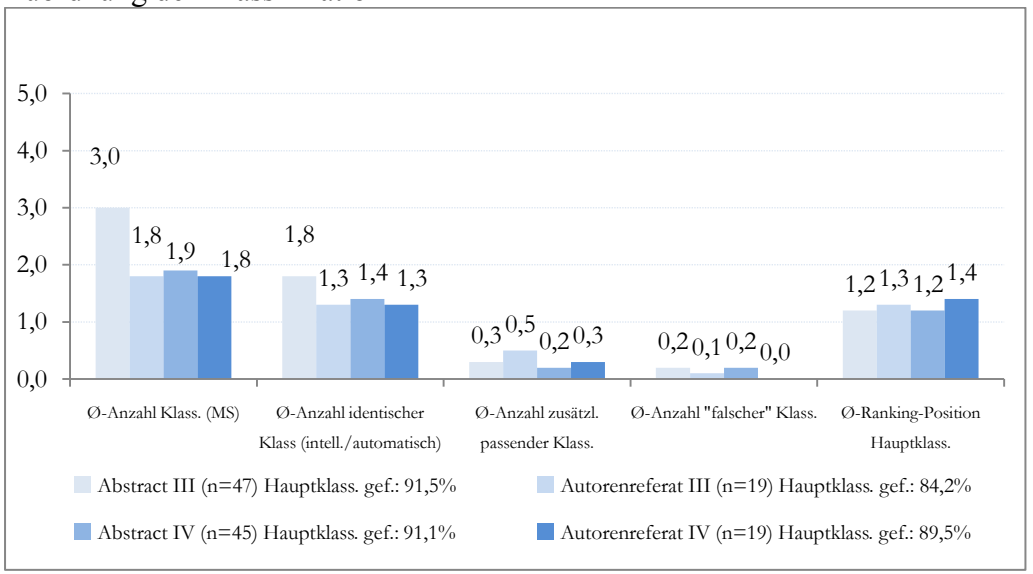


Politikwissenschaft

Vergabe der Deskriptoren

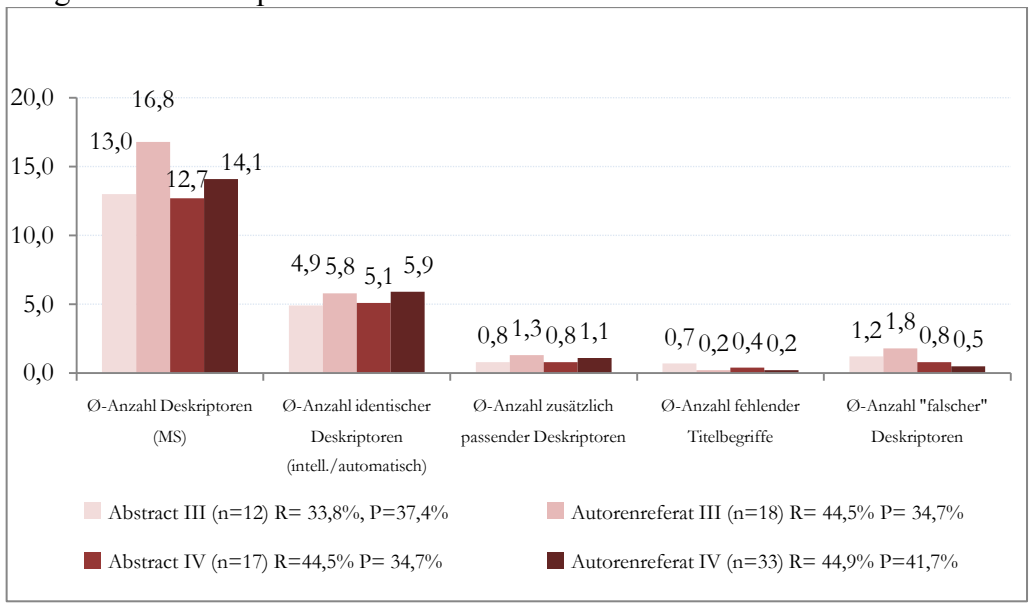


Zuordnung der Klassifikation



Geisteswissenschaften

Vergabe der Deskriptoren



Zuordnung der Klassifikation

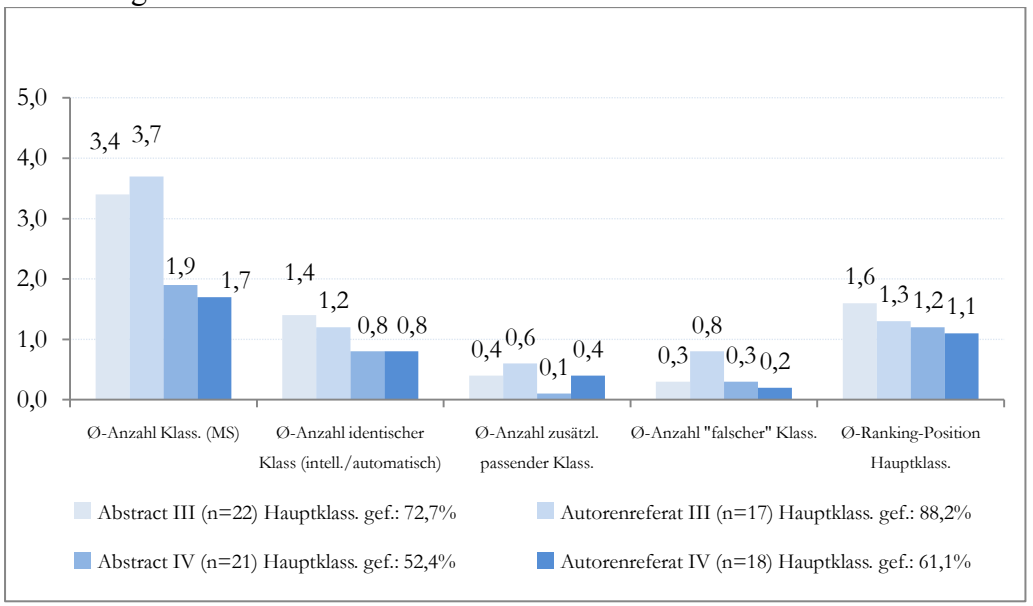
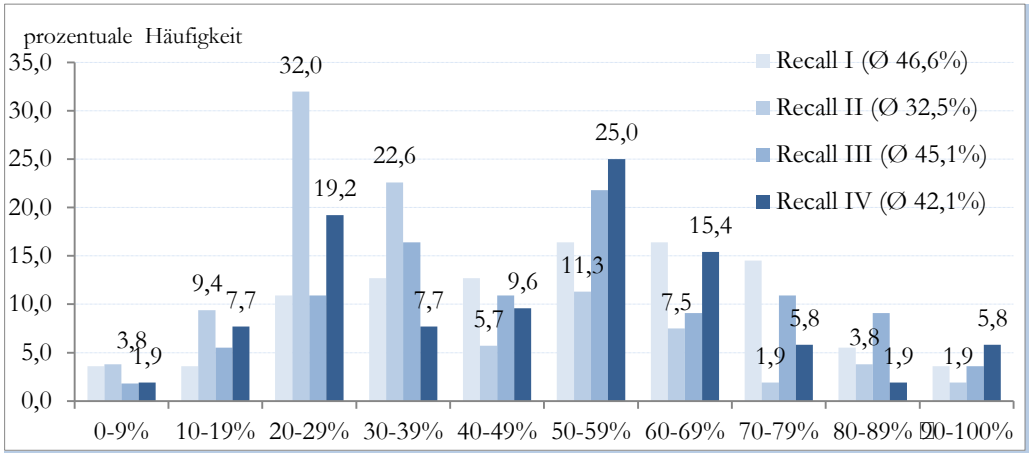


Tabelle 13: Vergleich der automatisch generierten Indexierungsergebnisse nach Abstract-Art.

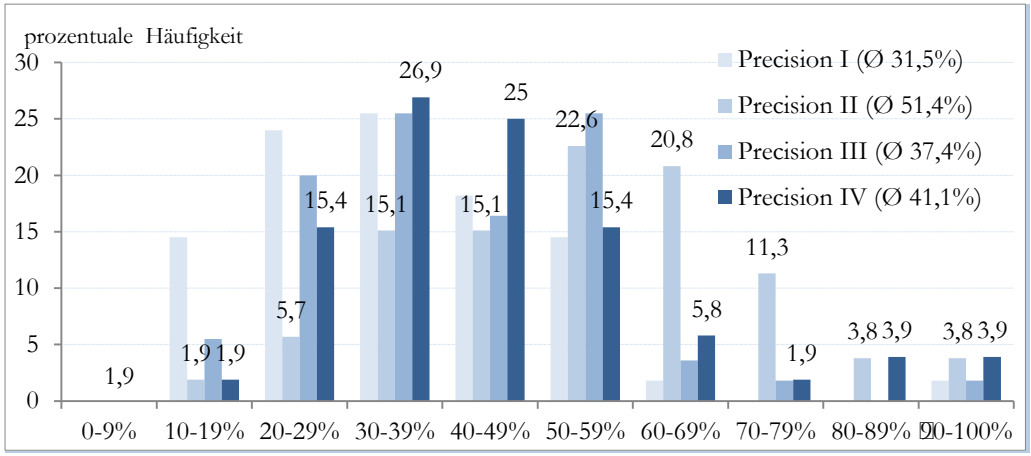
Übersicht über die *Recall*- und *Precision*-Werte der Fachteilgebiete Soziologie, Politik- und Geisteswissenschaften für die Testläufe I bis IV

Soziologie

Recall

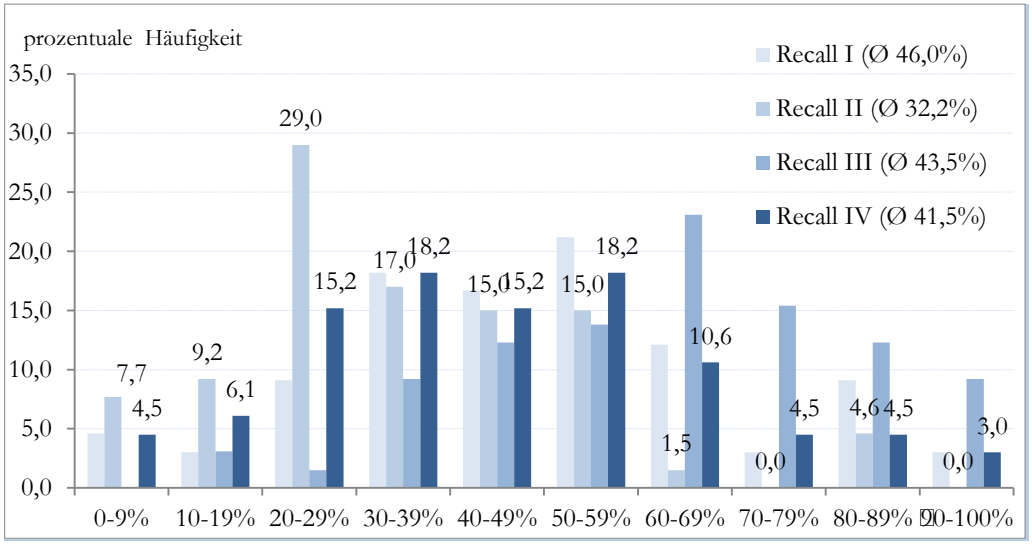


Precision

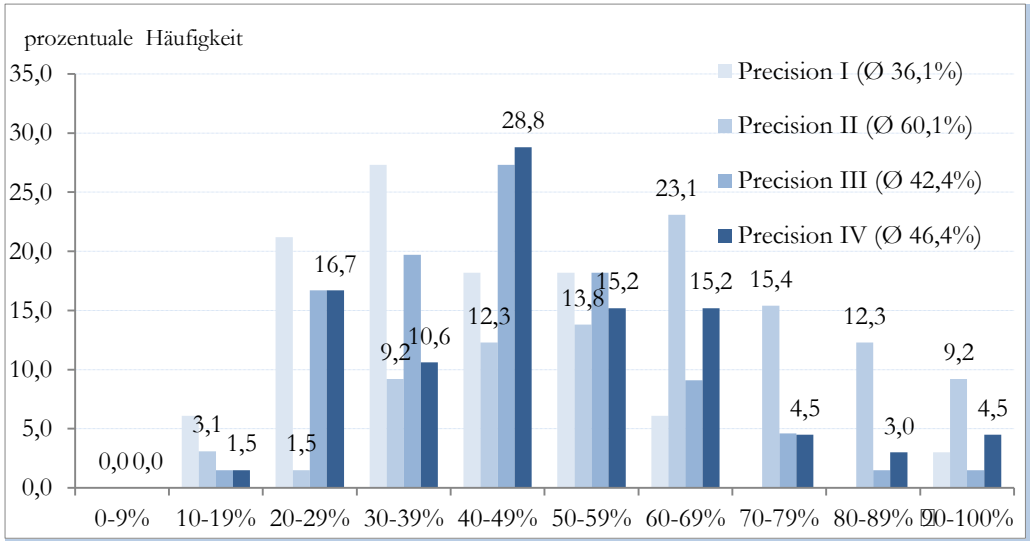


Politikwissenschaft

Recall

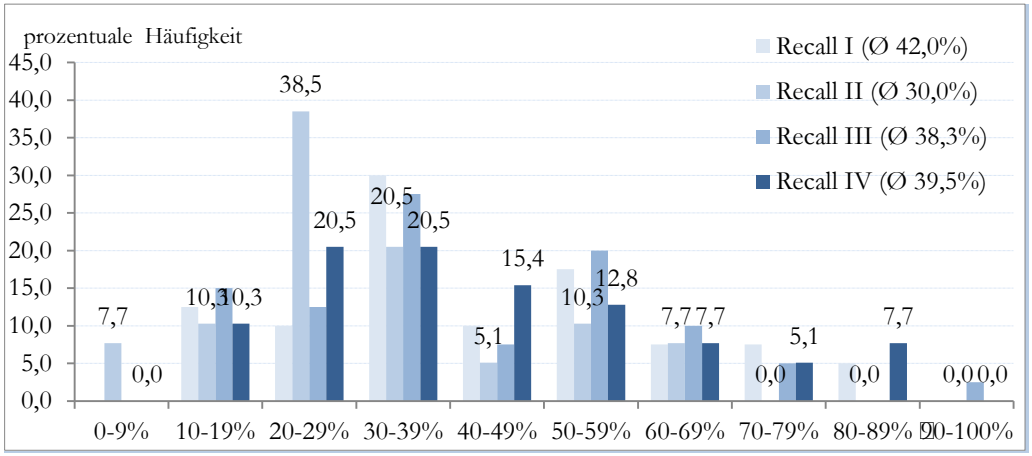


Precision



Geisteswissenschaften

Recall



Precision

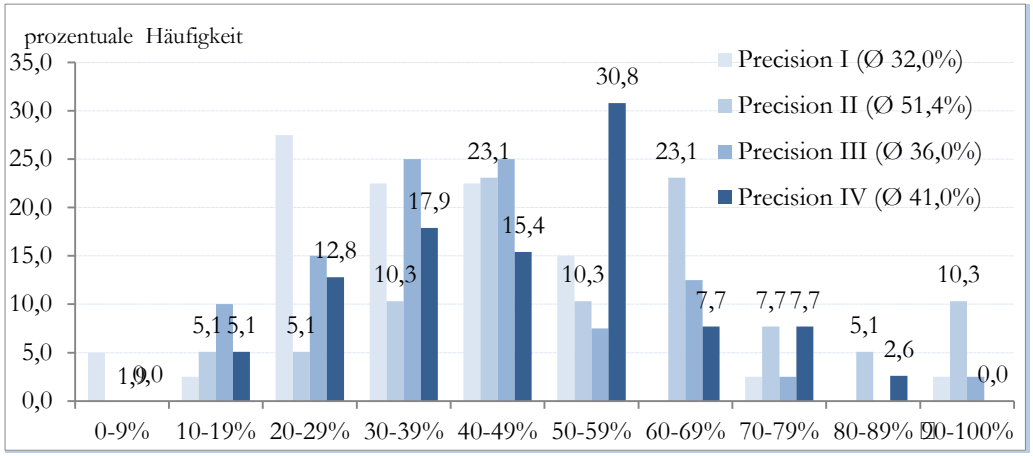


Tabelle 14: Übersicht der *Recall*- und *Precision*-Werte für die Fachteilgebiete Soziologie, Politik- und Geisteswissenschaften.

Auswertungstabelle für die Fachteilgebiete Soziologie, Politik- und Geisteswissenschaften

	Deskriptoren																			
	Ø-Anzahl					Ø identisch				Ø zusätzl. passend				Ø "falsche" Deskr.				Ø fehlende Titelbegr.			
	in-tell.	I	II	III	IV	I	II	III	IV	I	II	III	IV	I	II	III	IV	I	II	III	IV
Soziologie (n=53-56)	13,5	20,0	8,6	16,3	13,9	6,3	4,4	6,1	5,7	0,7	0,8	1,0	0,8	6,2	0,3	0,9	0,8	0,5	0,6	0,5	0,5
Politikw. (n=67-68)	16,2	20,6	8,7	16,6	14,4	7,4	5,2	7,0	6,8	0,4	0,8	1,6	1,2	5,2	0,4	0,5	0,6	0,3	0,5	0,3	0,2
Geistesw. (n=39-40)	13,8	18,1	8,1	14,7	13,3	5,8	4,2	5,3	10,7	0,5	0,7	1,0	0,9	3,5	0,7	1,5	0,7	0,6	0,6	0,5	0,3

	Recall (R)/*Precision* (P) (%)									Vorkommen in %											
	auf der Basis identischer Deskr.					inkl. zusätzl. pass. Deskr.				Ø zusätzl. passend				Ø "falsche" Deskr.				Ø fehlende Titelbegr.			
		I	II	III	IV	I	II	III	IV	I	II	III	IV	I	II	III	IV	I	II	III	IV
Soziologie	R	46,6	32,5	45,1	42,1	51,8	37,8	52,2	48,3	40,0	56,6	60,0	59,6	87,7	20,4	44,6	45,3	31,6	46,3	39,3	37,7
	P	31,5	51,4	37,4	41,1	35,1	62,7	43,4	47,1												
Politikw.	R	46,0	32,2	43,5	41,5	48,3	37,2	53,1	49,0	31,8	53,9	80,3	68,2	84,9	29,2	33,3	37,9	27,3	36,9	25,8	21,2
	P	36,1	60,1	42,4	46,4	48,3	37,2	53,1	49,0												
Geistesw.	R	42,0	30,0	38,3	40,0	45,6	34,9	45,8	45,6	35,0	43,6	62,5	61,5	77,5	38,5	57,5	43,6	42,5	51,3	37,5	23,1
	P	32,0	51,4	36,0	41,0	34,8	59,7	43,0	47,3												

	Klassifikation																							
	Ø-Anzahl					Ø identisch				Ø zusätzl. passende[148]				Ø "falsche" Klass.				Hauptklass. gef.(%)				Ø Ranking-Position			
	in-tell.	I	II	III	IV	I	II	III	IV	I	II	III	IV	I	II	III	IV	I	II	III	IV	I	II	III	IV
Soziologie (n=52-56)	2,3	4,3	3,6	3,9	1,9	1,3	1,3	1,3	0,9	0,2 19,3	0,6 44,2	0,7 54,6	0,5 43,4	2,0 76,4	0,5 41,2	0,4 37,0	0,2 21,2	78,2	78,4	85,2	69,2	1,7	1,6	1,4	1,2
Politikw. (n=65-68)	2,3	4,0	3,6	2,9	1,9	1,9	1,7	1,7	1,4	0,1 9,1	0,2 17,5	0,4 33,3	0,2 20,3	1,3 62,1	0,4 34,9	0,1 13,6	0,1 10,9	92,4	87,3	89,4	87,9	1,7	1,5	1,2	1,4
Geistesw. (n=38-40)	2,3	3,8	3,3	3,4	1,9	1,0	1,0	1,3	0,8	0,1 10,5	0,3 30,8	0,5 40,0	0,3 26,3	1,1 60,5	0,5 43,6	0,5 37,5	0,3 23,7	63,2	59,0	77,5	57,9	2,7	1,7	1,4	1,1

Tabelle 15: Allgemeine Auswertungstabelle für die Fachteilgebiete Soziologie, Politik- und Geisteswissenschaften Testläufe I bis IV

[148] Der untere Wert steht jeweils für das prozentuale Vorkommen zusätzlich passender bzw. inhaltlich fehlerhafter Klassifikationsvorschläge.

Abbildungsverzeichnis

Tabellenverzeichnis